UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

SPEECH-TO-SPEECH TRANSLATION TO SUPPORT MEDICAL INTERVIEWS

João António Santos Gomes Rodrigues

PROJETO

MESTRADO EM ENGENHARIA INFORMÁTICA
Especialização em Interacção e Conhecimento

2013

UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

SPEECH-TO-SPEECH TRANSLATION TO SUPPORT MEDICAL INTERVIEWS

João António Santos Gomes Rodrigues

PROJETO

MESTRADO EM ENGENHARIA INFORMÁTICA
Especialização em Interacção e Conhecimento

Projeto orientado pelo Prof. Doutor António Manuel Horta Branco
2013

Agradecimentos

Agradeço ao meu orientador, António Horta Branco, por conceder a oportunidade de realizar este trabalho e pela sua sagacidade e instrução oferecida para a concretização do mesmo.

Agradeço aos meus parceiros do grupo NLX, por terem aberto os caminhos e criado inúmeras ferramentas e conhecimento para o processamento de linguagem natural do português, das quais faço uso neste trabalho. Assim como pela imprescindível partilha e companhia.

Agradeço à minha família e à Joana, pelo apoio inefável.

Este trabalho foi apoiado pela Comissão Europeia no âmbito do projeto METANET4U (ECICTPSP:270893), pela Agência de Inovação no âmbito do projeto S4S (S4S / PLN) e pela Fundação para a Ciência e Tecnologia no âmbito do projeto DP4LT (PTDC / EEISII / 1940 / 2012). Às três expresso a minha sincera gratidão.

Resumo

Este relatório apresenta a criação de um sistema de tradução fala-para-fala. O sistema consiste na captação de voz na forma de sinal áudio que de seguida é interpretado, traduzido e sintetizado para voz. Tendo como entrada um enunciado numa linguagem de origem e como saída um enunciado numa linguagem destino.

O sistema implementado tem como âmbito do seu funcionamento o domínio médico, tendo em vista apoiar o diálogo entre médico e utente em linguagens diferentes durante consultas médicas. No caso do presente trabalho, foram escolhidos o português e inglês, sendo possível a tradução fala-para-fala nos dois sentidos. A escolha destas duas línguas resulta sobretudo da disponibilidade de recursos para o desenvolvimento do sistema.

Ao longo dos anos tem existido um esforço de pesquisa e desenvolvimento em tecnologia que permite quebrar as barreiras do multilinguismo. Uma dessas tecnologias, com resultados de qualidade crescentemente aceitável, são os sistemas de tradução fala-para-fala. Em geral, estes sistemas são compostos por três componentes: reconhecimento de fala, tradução automática e sintetização de voz. Neste projecto foram implementadas as três componentes. No entanto, uma vez que face às tecnologias disponíveis, a componente de tradução tem um maior impacto no desempenho final do sistema, a esta foi conferida uma maior atenção.

Embora nós, como humanos, compreendamos facilmente a linguagem falada, isto é algo extremamente difícil e complexo de um ponto de vista computacional. O objectivo do reconhecimento de fala é abordar esta tarefa computacionalmente através da construção de sistemas que mapeiam um sinal acústico para uma sequência de caracteres. Os modelos actuais para reconhecimento de fala fazem uso de modelos estatísticos. Nestes, a fala é reconhecida através do uso de modelos de linguagem que possibilitam a estimativa das probabilidades para as palavras, independentemente do sinal de entrada, e de um modelo acústico onde as propriedades acústicas da fala estão contempladas.

Os modelos actuais de tradução automática, assim como os de reconhecimento de fala, são na sua larga maioria estatísticos. Actualmente os modelos de tradução baseados em unidades frásicas de input são os que obtém os resultados com melhor qualidade. Esta abordagem consiste na tradução de pequenos segmentos de palavras, onde existe uma tradução lexical e um modelo de alinhamento. Os modelos estatísticos fazem uso de textos de duas línguas alinhados, tendo como princípio o facto de que através da frequência

de cada segmento de palavras, em relação à outra linguagem, seja obtida uma distribuição probabilística. Deste modo torna-se possível calcular qual a palavra ou conjunto de palavras mais prováveis de ocorrer como tradução para determinado texto que se pretenda traduzir.

A sintetização de voz consiste na geração de fala na forma de onda acústica tendo como ponto de partida uma palavra ou uma sequência de palavras. Envolve o processamento de linguagens naturais e processamento de sinal. O primeiro converte o texto numa representação fonética e o último converte essa representação em sinal acústico.

Neste documento é apresentado o estado da arte das três áreas envolvidas. São também apresentados os sistemas de tradução fala-para-fala, fazendo ou não uso do domínio médico, e também os processos existentes para a avaliação de cada uma das componentes.

Tendo em vista a implementação de um sistema com as diversas componentes, foi necessário efectuar um levantamento da tecnologia existente. O levantamento teve por objectivo a implementação de duas soluções aplicacionais. Uma aplicação disponível pela internet como página web e outra através de uma aplicação móvel, ambas permitindo o reconhecimento de fala, tradução automática e sintetização de voz em ambas as linguagens e direcções.

Dois sistemas de reconhecimento de fala foram escolhidos, o Microsoft Speech Platform para a aplicação móvel e o reconhecimento de fala disponível pelo Google nos browsers Google Chrome. O primeiro a ser usado na aplicação móvel e o segundo na aplicação web. O sistema de tradução automática escolhido foi o Moses. Sendo um sistema de tradução estatístico que permite a criação de modelos de tradução diversos, como os modelos baseados em frase e os modelos baseados em fatores. O sistema de sintetização de voz escolhido foi o Microsoft Speech Platform.

A aplicação móvel foi desenvolvida para a plataforma iOS da Apple tendo em vista o uso de um telemóvel iPhone. A integração dos componentes pelas diversas arquitecturas foi assegurada pela implementação de web services.

O reconhecimento de fala na aplicação web foi desenvolvido recorrendo ao uso da W3C Speech Input API Specifications, onde a programação através de HTML permite a captação de áudio no Google Chrome.

Para a implementação do sistema tradução fala-para-fala foi necessário a obtenção de corpora paralelos de forma a se poder treinar os modelos estatísticos, sendo este um dos factores cruciais para o bom desempenho dos componentes. Uma vez que o sistema tem como domínio de aplicação o diálogo médico, corpora neste domínio seria o mais vantajoso. No entanto, a inexistência de tais corpora para o par Inglês-Português levou à aquisição de corpora alternativos.

Através de uma experiência exploratória foi abordado o tipo de implementação mais adequado da componente de reconhecimento de fala, tendo como foco o modelo de linguagem. Três experiências foram então conduzidas de forma a decidir entre a aplicação

de uma modelo de linguagem baseado em regras ou um modelo estatístico.

Para implementar um modelo de linguagem baseado em regras foi necessário a criação de um corpus médico que reflectisse um diálogo entre médico e paciente. Para tal, com a ajuda de um médico, criei um diálogo de um caso hipotético de lesão num braço devido a um acidente de carro. Este diálogo teve como base para a sua estruturação a aplicação do processo de anamnesis. A anamnesis consiste numa metodologia médica que através de um conjunto de perguntas chave permite adquirir a informação necessária para a formulação de um diagnóstico médico e decisão sobre o tratamento necessário. O corpus médico foi também transformado num corpus de fala de forma a este ser avaliado ao longo das experiências.

Numa primeira experiência foi criada uma gramática básica cuja implementação foi obtida recorrendo à Speech Recognition Grammar Specification de forma a ser usada como modelo de linguagem pela componente de reconhecimento de fala.

A segunda experiência tinha como objectivo a criação de uma gramática mais complexa que a primeira. Para tal foi criada uma gramática livre de contexto. Após a criação da gramática livre de contexto esta foi convertida manualmente para uma gramática SRGS.

Na terceira experiência foram criados dois modelo de linguagem estatísticos, o primeiro fazendo uso do mesmo corpus que o usado nas experiências anteriores e o segundo composto por 30.000 frases independentes.

Obteve-se melhores resultados com o modelo de linguagem estatístico e este ficou como a escolha para a implementação do componente de reconhecimento de fala.

No treino da componente de tradução automática foram usados dois modelos estatísticos, baseados em frases e em factores. Pretendeu-se comparar os resultados entre os dois modelos de forma a escolher o modelo mais vantajoso. Para fazer uso do modelo baseado em factores foi necessária a preparação de corpora. Com os corpora já adquiridos foi concretizada a sua anotação para ambas as linguagens. Recorrendo ao LX-Suite e ao CoreNLP, foram criados corpora anotados com lemmas e informação morfosintática, com a primeira ferramenta para o português e a última para o inglês.

Uma vez que a componente de sintetização de voz permitia uma implementação célere, esta foi implementada recorrendo aos modelos já existentes para ambas as linguagens e disponibilizados pela ferramenta.

Por fim, são apresentados os resultados obtidos e a sua avaliação. Tanto a avaliação do sistema de reconhecimento de fala como o de tradução automática demonstraram um desempenho muito competitivo, do nível do estado da arte.

A componente de reconhecimento de fala, assim como a componente de tradução automática, obtiveram melhores resultados fazendo-se uso de modelos de linguagem estatístico.

Palavras-chave: Tradução fala-para-fala, Reconhecimento de fala, Tradução Automática Estatística, Sintetização de voz, Domínio Médico

Abstract

This report presents the development of a speech-to-speech translation system. The system consists in the capture of voice as an audio signal that is then interpreted, translated and synthesized to voice for a target language.

The three main components of the system, speech recognition, machine translation and speech synthesis, make use of statistical models, such as hidden Markov models. Given the technology available, the machine translation component has a greater impact on the performance of the system, a greater attention has thus been given to it.

The system assumes the support to medical interviews between doctor and patient in different languages as its applicational domain.

Two application solutions were developed: an online service on a website and a mobile application.

This report begins by presenting the general concepts of the relevant areas involved. It proceeds with an overview of the state of the art relating to each area as well as to the methods used for the evaluation of the different components. It provides also an overview of existing technology and the criteria for choosing the tools to be used in the development of the system. It explains the acquisition and creation of the corpora used, and the process of development and integration of the components: speech recognition, machine translation and text-to-speech.

Finally, the evaluation results are presented, as well as the final conclusions.

Palavras-chave: Speech-to-Speech Translation, Automatic Speech Recognition, Statistical Machine Translation, Speech Synthesis, Medical Domain

Contents

List of Figures

List of Tables

Chapter 1

Overview

This document reports on the work done within the course of Informatics Engineering Project (PEI) of the MSc in Informatics Engineering, of the Faculty of Sciences, University of Lisbon. This project was developed in the research unit NLX-Group (Natural Language and Speech Group) of the Department of Informatics. It was integrated in the European R&D project METANET4U (ECICTPSP:270893), the Agência de Inovação project S4S (S4S / PLN) and in the FCT project DP4LT (PTDC / EEISII / 1940 / 2012), coordinated by Prof. António Branco.

1.1 Introduction

It is estimated that exist more than 6900 living languages [2]. Coping with this "Tower of Babel" has always been one of mankind's needs. Over the years, technologies have been researched and developed that help to overcome the barriers of multilingualism. An important contribution to this endeavour, allowing a result with acceptable quality, is Speech-to-Speech Translation (S2ST).

S2ST is the translation of utterances in a source language to utterances in a target language, where the source language is received in the form of natural speech and the target language is output by synthesization of speech. Its use can be found in many applications such as: interpretation, phone communication, tourism information or film subtitling, etc.

S2ST may consist of three sequential components: Speech Recognition, Machine Translation and Text-to-Speech Synthesis. From each one of them result distinct areas of research in Natural Language Processing. Before presenting an overview of S2ST, each component is thus addressed in isolation.

1.1.1 Machine Translation

Machine Translation (MT) refers to computerized systems responsible for the production of translations, from a source language to a target language, with or without human as-

sistance. First proposed in 1947 by Warren Weaver, Norbert Wiener and A. D. Boot, it lead to some substantial research in United States, Soviet Union and Western Europe at the time. By 1966, the publication of a report by ALPAC, concluded that there was little prospect of quality/cost effective machine translations and many of the on going projects were stalled by lack of investment. By then the translation was done through a method called direct translation. This method used few syntactic or semantic analysis, simply employing a database of words and their translations. This solution had several deficiencies. A famous example is the machine translation, from English to Russian, of the sentence "The spirit is willing, but the flesh is weak" in English, resulting in the Russian equivalent to "The vodka is good, but the meat is rotten".

This bring us to the main difficulty of MT, coping with the ambiguity of natural language. Ambiguity can arise in different ways: lexical, where a single word can have more than one sense, hence, more than one translation. Or in a structural way, where the sequence of words can be interpreted in different ways. As an example of lexical ambiguity, the Portuguese word *bateria* can have the meaning of a set of music instruments or of a battery. One finds an example of structural ambiguity in the sentence "John saw Mary with the binoculars", which can be interpreted as John seeing Mary through binoculars or John seeing Mary in the possession of binoculars. It will be necessary to decide which meaning is intended so the correct translation can be obtained.

The 1970s and 1980s brought the incorporation of more sophisticated linguistic and computational techniques to the development of the so-called second generation of MT design. This is characterized in particular by the indirect approach to translation, in which the source text is transformed into the corresponding target text via an intermediate, linguistically motivated, representation, by using rules (rule-based method) for morphology, syntax, lexical selection, semantic analysis and generation.

Interlingua was one of these approaches, a theoretically "pure" one. The representation used is an abstract representation of the meaning of the source text, capturing all and only the linguistic information necessary to generate an appropriate target text showing no undue influence from the original text. It assumes that it is possible the conversion from a source language into a semantic-syntactic representation which is common to more than one language. Although it would represent an extreme economy of effort, taking into account the large number of languages it would be possible to translate, the complexity of these approach is extremely high. This revealed however to be unattainable in practice. The problem of representing meaning in a formal way is one of the grand challenges of artificial intelligence, with interesting philosophical implications.

A more practical approach was made with the so-called Transfer approach. This translation method is made in a three step process:

- Analysis of the source language input into a syntactic structure representation. In later forms of transfer systems, this analysis step involves the delivering of some

kind of semantic representation.

- Transfer of that intermediate representation of the source language into the corresponding intermediate representation of the target language.
- Synthesis of the output in the target language structure from the latter intermediate representation.

This approach brought some of the early commercially successful systems, such as Météo, Systran, Logos and METAL, which are still in operation today.

Another approach developed for the MT is the so-called example-based translation. This type of system tries to find a sentence similar to the input sentence in a parallel corpus, a large collection of sentences aligned with its translations, and delivers the respective translation that may be found in that collection.

In the late 1980s, IBM published a ground-breaking research paper [3] in Statistical Machine Translation (SMT) inspired by the success that the statistical approach had with speech recognition. While research on statistical methods for machine translation continued throughout the 1990s, the approach gathered full steam only around the year 2000.

Crucial factors contributed to the SMT thrive. The United States of America invested in the SMT research with the DARPA foundation of TIDES and GALE programs. Another reason for the investment was due to the terrorist attack of September 11, 2001, given the importance for the automatic translation of foreign languages, especially Arabic.

SMT is now the dominant approach in the field and applied in the translation technologies, such as the ones used by Google and Microsoft .

One of the first methods applied in SMT was the word-based translation method. In this method the translation is based on lexical translation and on an alignment model. Using a dictionary that maps the words of the two languages, the frequency of each word in the parallel aligned corpus used for training is counted in relation to the translation result. With these values, one obtains a probability distribution for each word forming a lexical translation. Since the order of words differ between languages, the method takes into account the reordering during translation, using an alignment function, which maps the position of the word between source and target language.

More recently various works have shown better translation quality with the use of a phrase-based translation method. This approach translates small word sequences at a time, and extends the previous translation model and reordering model of word-based translation by applying it to phrases. There have also been research with extensions to the phrase model, that use syntactic analysis with tree-based models and also hybrid rule-based and statistical methods[4][5][6][7].

The ideal situation would be to produce quality translations without human intervention, but currently this is not possible yet except in highly constraint situations. Such constrained situation uses a limited domain or a controlled language, called a sublanguage, that restricts the vocabulary and the grammatical structures of the input texts, reducing the problems of ambiguity and selection of equivalents.

1.1.2 Automatic Speech Recognition

Although we, as humans, easily understand spoken language, this is another remarkably difficult and complex task from a computational point of view.

The goal of automatic speech recognition (ASR) is to address this complex task by computational means, through the construction of systems that map an acoustic signal into a string of words. In recent years, ASR technology matured to the point where it is viable in certain application domains. We can find various applications, from human-computer interaction, telephony and dictation, to systems providing updated travel information, stock price quotations or weather reports, as it can be found in Apple's Siri.

Speech recognition research started more than 70 years ago. It began in the 1930s using ad-hoc methods to recognize sounds or isolated words. From the 1950s to 1960s, a second generation of systems started to be researched using acoustic-phonetic approaches to recognize phonemes, phones, or digit vocabularies.

The third generation, 1960s to 1980s, used pattern recognition approaches with vocabularies with a small and medium size, with isolated and connected word sequences. It applied Linear Predictive Coding[1] for spectral representation, pattern clustering methods for speaker-independent systems, and dynamic programming methods for solving connected word recognition problems. The fourth generation research, 1980s to 2000s, went into tackle large vocabulary (more than a thousand words) using Hidden Markov models (HMM), statistical methods for modeling speech dynamics and continuous speech recognition systems. Various performance criteria and methods were used to optimize statistical models like neural networks. The present generation, the fifth, is seeing the use of parallel processing methods to increase recognition decision reliability by doing so with combinations of HMMs and acoustic–phonetic approaches to detect and correct linguistic irregularities. There have been research also in the robustness for recognition of speech in increased ambient noise.

As mentioned, the current systems for ASR use statistical models. The speech is recognized using a language model which provides the estimates of probabilities for all

[1]Linear predictive coding is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model.

words independently of the input signal, and an acoustic model, where the acoustic properties are contained. The main components of a generic speech recognition system are:

- Acoustic parameterization and modeling takes into account the channel, noise, speaker gender and independence of speaker.
- Lexical and pronunciation modeling bridges the acoustic-level representation and the word sequence output by the speech recognizer, involving two main parts: definition and selection of vocabulary items and representation of each pronunciation entry using basic acoustic units of the recognizer.
- Language Modeling is used to estimate the probability of word sequences. Grammatical constraints can be described using a context-free-grammar[2] or modeled stochastically, being the n-gram[3] the more popular model.
- Decoding is the search algorithm for the combination of the acoustic model and the language model. It determines the word sequence with the highest likelihood given the acoustic, the lexical and the language models.

1.1.3 Text-to-Speech Synthesis

Text-to-Speech Synthesis (TTS) is the task of generating speech, as acoustic waveforms, from text. Its application can be found from telephone based conversational agents, to help for reading by the blind, and includes support to video games, toys, car navigation or to replace the actual voice for patients of neurological disorders.

The mapping of TTS occurs in two stages: natural language processing and digital signal processing. The former converts text into an internal phonemic representation and the latter converting the internal representation into an acoustic wave form. The natural language processing module contains:

- Preprocessor which deals with the normalization of text, including numbers, abbreviations, acronyms and idiomatic expressions, and with the storing of each word in an internal data structure.
- Morphosyntactic analyzer which provides the segmentation of text into basic linguistic entities, the grammatical structure of sentence and the syntactic characteristics of each word.
- Syntactic-prosodic parser, which finds the hierarchical organization of words into phrase like constituents that more closely relates to its expected intonational structure and also provides contextual analysis.

[2] A context-free grammar is a set of recursive rewriting rules (or productions) used to generate patterns of strings.

[3] An n-gram is a contiguous sequence of n items from a given sequence of text or speech.

- Phonetizer (or letter-to-sound), which automatically determines the phonetic transcription of the incoming text, and takes care of the pronunciation and rhythmic constraints. It can be implemented with dictionary-based or rule-based strategies. The former stores a maximum of phonological knowledge into a lexicon, the latter transfers most of the phonological competence of dictionaries into a set of letter-to-sound rules.

- Prosody generator, which takes care of the composition of speech signal properties: pitch, loudness and syllable length. It allows to focus on certain syllables and, on the use of intonation.

The digital signal processing module, in turn, employs two main types of techniques for speech generation: synthesis by rule or concatenative synthesis.

The rule based methods are formant-based or articulation based. The former uses additive synthesis[4] and an acoustic model without voice sample. The latter models the mechanical motions of the articulators and the resulting distributions of volume velocity and sound pressure in the lungs, larynx, vocal and nasal tracks.

The second main technique is the concatenative synthesis, in which samples of speech that were chopped up are stored in a database, and then combined and reconfigured to create new spoken utterances.

Nowadays most modern commercial TTS are based on concatenative techniques.

1.1.4 Speech-to-Speech Translation

The integration of the three processes just described, Speech Recognition, Machine Translation and Text-to-Speech Synthesis, can simply be done by using the output from each component as input to the next. Consequently when one of the first two component delivers results with errors, these are propagated, affecting the performance of the whole system. So along with the need for robust individual components, the S2ST should be error-tolerant.

One strategy to cope with this requirement is to use multiple hypotheses from the speech recognition, replacing the string output by a word lattice[5] or n-best list.

The machine translation can also provide multiple interpretation results helping the TTS to obtain better results.

[4] Additive synthesis is a sound synthesis technique that creates timbre by adding sine waves together.

[5] A word lattice is a directed acyclic graph with a single start point and edges labeled with a word and weight, representing any finite set of strings.

1.2 Objectives

The main objective of the work described in the present document was to create a Portuguese and English bi-directional S2ST prototype system motivated by an applicational context of doctor-patient communication in medical interviews.

The S2ST system needed to resort to all the three components, Speech Recognition, Machine Translation and Text-to-Speech Synthesis. The interface implementation was done in two prototype solutions, web and mobile.

It was conceived to be suitable for a communication situation where the physician and patient need to communicate with each other and they do not master a common natural language.

1.3 Planning

The work reported here was planned along different tasks, as displayed in detail in Table 1.1.

1.4 Organization of the Document

The present report began by introducing the general concepts involved in S2ST, providing an overview of the main components. This introduction was followed by the indication of the objectives, of the planning of work and of the organization of this document (in the present section).

This document proceeds with providing an overview of the state of the art for each component of the S2ST and also with the discussion of the evaluation and domain criteria and methodologies.

It continues by providing an overview of existing technology and the criteria for choosing the application support tools I have adopted. It reports on the acquisition and creation of the parallel corpora used for training the system, and on the development process and the integration of the different components.

Finally, the results are discussed and a conclusion is presented.

- **Chapter 1** - (The present chapter). Introduction to the main components: Machine Translation, Automatic Speech Recognition and Text-to-Speech Synthesis tools and the combination of the three into the S2ST system. The objectives for the present work are defined, and the planning schedule is detailed.

- **Chapter 2** - Presentation of the state of the art, the methods and approaches adopted for each one of the three components, along with their justification. The methods for evaluation of each component and an overview of the medical domain of application is also presented.

- **Chapter 3** - Description of available technology and the criteria used for the selection of the tools used to create my S2ST system.
- **Chapter 4** - Discussion of the implementation of the system.
- **Chapter 5** - Evaluation of the implemented S2ST system.
- **Chapter 6** - Conclusions and future work.

Task	Due date
1) General study of each of the components and their integration (automatic speech recognition, machine translation and speech synthesis) by resorting to the relevant literature.	Middle September 2012
2) Writing of the overview along with the objectives.	Late September 2012
3) Search on and summarization of the state of the art.	Middle October 2012
4) Surveying on the technology available for the development of the system along with the design criteria opted for.	Late October 2012
5) Preparation and creation of corpora. Gathering of parallel corpora and preparation for each component. Defining a language model, exploring the statistical approach and the context-free grammar for the ASR System taking a real life medical diagnosis as a basis. Creating an annotated corpus in both languages for the factored approach of statistical machine translation.	Late December 2012
6) Development of the solutions with the selected tools. Setting up of each of their components and their integration. Development of a web services interface for each component. Development of web and mobile application taking into account human-computer interaction.	Late February 2013
7) Evaluating the system and proceed with adjustments if necessary.	Middle March 2013
8) Global assessment of the results.	Middle April 2013
9) Gathering all the information and writing the project report.	Late June 2013

Table 1.1: Planning schedule

Chapter 2

State of the Art and Related Work

2.1 Machine Translation

As mentioned in the first chapter, several approaches can be taken for the creation of a machine translation system. In terms of state of the art, the statistical approach is currently the most popular one, and in general competitive results can be obtained using the Statistical Phrase-Based Model [4].

However, the rule-based systems can still be found in some industry solutions and there are research being done with rule-based methods. These methods are quite useful when one lacks training data or when they are used for a limited domain [8].

More recently, there have been an exploration of hybrid systems by integrating rule-based syntactic and semantic information with statistical machine translation phrase-based models [9][10].

2.1.1 Statistical Approach

The statistical approach for MT defines the optimal translation as the output sentence that has the highest probability of being the translation of a source sentence. The main challenges here consist in estimating the probability of a translation and in efficiently finding the sentence with the highest probability.

Word-based model

Word-based models were the first models that materialized a statistical machine translation approach. They are the original work of IBM Research project Candide [11]. It is not the state of art any longer but contains the principles and methods that are still used in current research.

Words can have multiple senses, being thus ambiguous. When we pick a word from a bilingual dictionary, we can find different possible translations. As an example, the Portuguese word *bateria* can have the following English translations:

Adeus		**azul**		**de**		**segunda-feira**	
t	$p(t\|s)$	*t*	$p(t\|s)$	*t*	$p(t\|s)$	*t*	$p(t\|s)$
goodbye	0.8	blue	0.95	of	0.6	monday	0.98
bye	0.1	azure	0.04	to	0.2	second-fair	0.02
farewell	0.07	sad	0.005	from	0.1		
bye-bye	0.03	melancholy	0.005	in	0.1		

Table 2.1: Example of a probability distribution for the source sentence *Adeus azul de segunda-feira*

Bateria - battery, drums, percussion, storage battery, accumulator.

Some of these translations are more frequent than others and some may be used in very specific contexts. Using a parallel corpus[1] as the training data, the word-based method counts how often each source word is translated into target words. Given the frequencies for each word, a probability distribution is then obtained [4][12]. The probability distribution is also called maximum likelihood estimation or translation probability, mathematically represented as $p(t|s)$, where *s* denotes a source word and *t* a target word. See table 2.1 for an illustration.

Translated words may need to be suppressed or/and reordered. This implicit need is solved by using an alignment model that maps the position from the target word to the position of a source word in the resulting sentence. The model works without knowing which of the words in the corpus is its translation. Since the word-alignment information does not exist in a parallel corpus, the method needs to work with incomplete data. The expectation maximization algorithm addresses this situation of incomplete data. On the basis of the probability of word pairings, it estimates the maximum likelihood of parameters, in this case the alignment model being the parameter.

Some translations are more appropriate than others, that is they can be more fluent, etc. Ensuring a fluent output is done by using a language model. A language model represents the probability of a string $p_{LM}(s)$ to belong to a given language. The most common method used to represent the language model is the n-gram language model.

To obtain the best translation, the language model and a translation model are combined:

$$argmax_t p(t|s) = argmax_t p(s|t)p_{LM}(t) \tag{2.1}$$

[1] A parallel corpus is a text placed alongside its translation.

The best translation is obtained from the largest probability of t given s. The $p(s|t)$ represents the translation model probability and the $p_{LM}(t)$ represents the language model probability of the a target sentence t. This prevents the occurrence of a strict word to word translation by taking into account the fluency.

Phrase-based model

The phrase-based model uses phrases instead of words as the smallest units for the translation. The principal benefits over the word-based method are [4]:

- Better handling of one-to-many mappings of the words.
- More precision on resolving translation ambiguities.
- When using a large training corpora, it is possible to learn longer and useful phrases, sometimes even memorize the translation of entire sentences.

The phrase-based model uses the same principles of the word-based model, in Equation 2.1, with the reformulation of the translation probability $p(s|t)$, as in the Equation 2.3, and with an additional factor, the word cost w, allowing the calibration of the output length. The w value is defined as a constant, so one can use a larger than 1 value for a longer output or a smaller than 1, biasing a shorter output.

$$argmax_t p(t|s) = argmax_t p(s|t) p_{LM}(t) w^{length(t)} \tag{2.2}$$

$$p(s|t) = p(\bar{s}_1^I|\bar{t}_1^I) = \prod_{i=1}^{I} \phi(\bar{s}_i|\bar{t}_i) d(start_i - end_{i-1}) \tag{2.3}$$

Each source phrase $\bar{s}_i$ from the input sentence $\bar{s}_1^I$ is translated into an target phrase $\bar{t}_i$ along with a probability distribution $\phi(\bar{s}_i|\bar{t}_i)$. The reordering is undertaken by a distortion probability distribution $d(start_i - end_{i-1})$, where $start_i$ is the start position of the source phrase that was translated into the ith target phrase, and end_{i-1} is the end position of the source phrase that was translated into the $(i-1)$th target phrase. The distortion probability is used to penalizes large distortions by giving lower and lower probability for larger distortions [13].

To learn the phrase translation table $p(s|t)$, a word alignment is created between each sentence pair of the parallel corpus and then extracted the phrase pairs that are consistent.

There are a number of possible extensions to the translation model that allows better translations, such as:

- Log-Linear Models, introducing weights that scale the influence of each component.
- Bidirectional Translation Probabilities, using the phrase translation probabilities in the actual translation direction.
- Lexical Weighting, a smoothing method.

In the recent years there have been several research works on how to integrate additional linguistic markup at the word level. The markup consists of the annotation of each phrase with the structure of the language, such as parts of speech or word function [14][15][10]. One of these approaches is the so-called Factored Translation Model.

The reasons for the integration of linguistic information are: a) the better overcoming of the data sparseness caused by limited training data, and b) bringing the already proved techniques (using rule-based methods) of syntactic, morphology or semantic into translation. The model uses a word as a vector of factors that can represent different levels of annotation. It can perform the morphological analysis and generation which consists of three steps: Translation of Lemmas, [2] translation of Part-of-speech and morphological information, and finally the generation of surface forms. See Figure 2.1 for an illustration, taken from Moses User Manual and Code Guide.[3] One of the main difference from phrase-based systems lies in the preparation of the training data, that needs to be annotated.

Decoding

Finding the best translation according to the formula chosen for a statistical machine translation system is done through so-called decoding. Heuristic search methods have been used to address this problem, a problem proven to be NP-complete.

Before the decoding can take place, a number of possible translations are collected, called the translation options. These are stored with the first target word covered, last target word covered, target phrase translation and phrase translation probability. The search is formulated as a succession of partial translations and hypotheses, which are expanded to construct the total translation. There are various search heuristics that can be used for this task, such as: Beam search based on coverage stack, A* Search or Greedy hill climbing. The most generic decoders search for a translation that maximizes a log-linear

[2] Lemmas are the canonical form of a set of words. For example, run, runs, ran and running are forms of the same lexeme, with run as the Lemma.

[3] http://www.statmt.org/moses/manual/manual.pdf - Last visited in 24/10/12

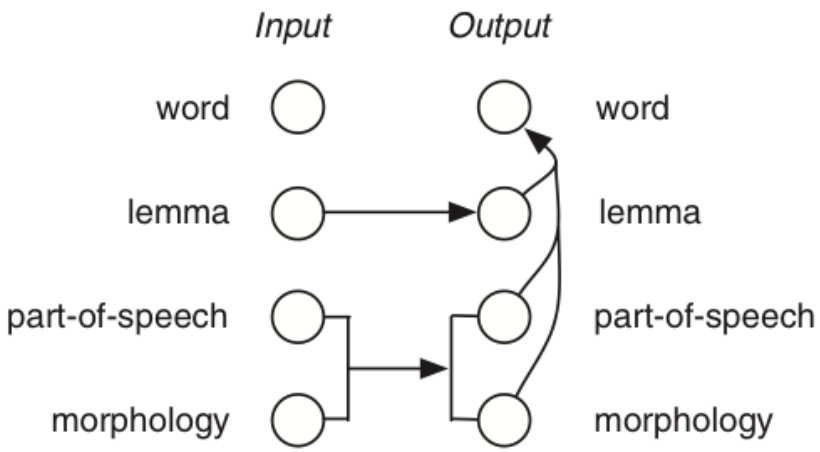

Figure 2.1: Factored Model mapping steps

weighted objective function which is defined over simple cases contained in a language model and a translation model [16].

Training Data

One of the principal concerns in SMT is the data used for training the models. As a general rule, more data tends to yield better language models [17]. Over the years, there has been a growing in the size of the collection of bilingual data and parallel corpus. Take for example the Europarl [18], a parallel corpus with over 11 languages from the proceedings of the European Parliament; or the JRC-Acquis, with more than 20 languages, containing European Union documents of mostly legal nature [19]. These parallel corpora tend to be in the order of hundreds of millions of words.

The data needs preparation before it can be used, so it passes over the processes of: sentence alignment, text de-formatting, tokenization, true-casing, cleaning and encoding [16]. There has been a significant amount of research for finding an alternative and viable way to extract parallel segments from non-parallel corpora, using techniques as maximum entropy classifier [20].

2.1.2 Rule-Based Approach

Although the rule-based is still used in many research works, the statistical approaches became more popular nowadays. The purpose of this section is to give a general idea of the rule-based approach so that some related work can be understood.

In the rule-based approach, the grammar is used and forms a theory of the grammatical

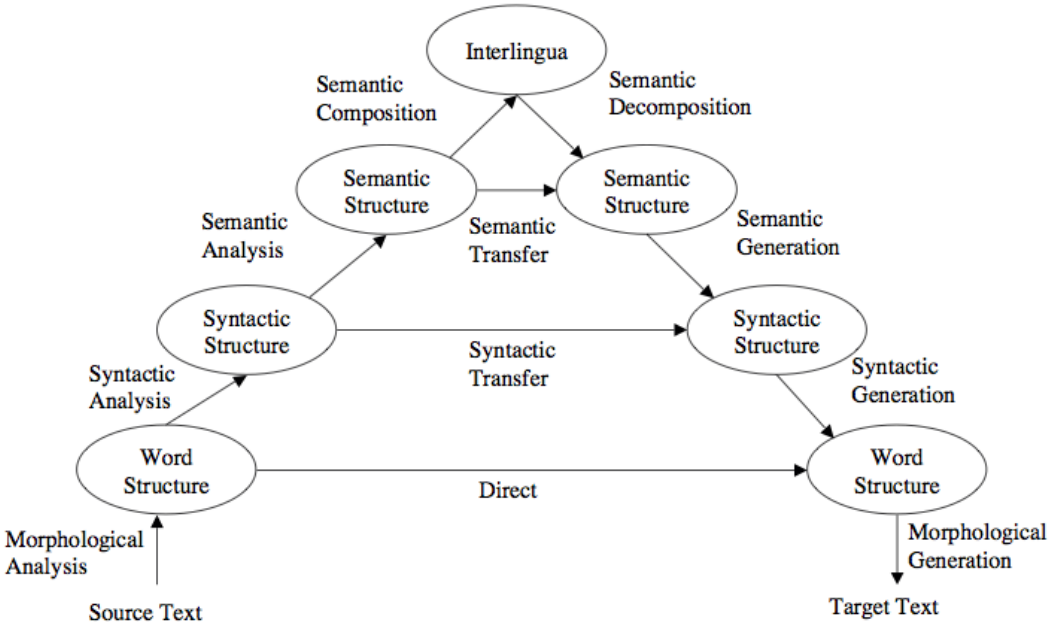

Figure 2.2: Vauquois Triangle

structure. By defining grammar rules, it is possible to generate or parse an infinite set of grammatical sentences. Usually, these rules are created based on a corpus. A knowledge based is defined, consisting of entity-oriented grammar formalism describing the semantic content of the language, and functional grammar formalism describing the syntax. The process consists of parsing the source text, extracting the meaning and encoding it into a semantic meaning representation.

There are different methodologies within the rule-based approach, as can be seen with the Vauquois Triangle in Figure 2.2, taken from [21]. Each methodology have a different depth of analysis:

- Direct transfer: uses simple rules at the word level.
- Syntactic and Semantic: creates an internal representation, for source and target languages, of the structure and meaning that is used to generate the translation.
- Interlingua: a unique internal representation is used for all source and target languages.

Usually for the Interlingua representation scheme, a second-order logic or a event-based representation is used, see Figure 2.3 for an illustration of the latter, taken from [6]. The large number of concepts that need to be represented requires a exhaustive analysis of semantics of the domain. As this is feasible in a relatively simple domain, interlingua is generally only used in a sublanguage domain. [6]

EVENT	SLAPPING		
AGENT	MARY		
TENSE	PAST		
POLARITY	NEGATIVE		
	WITCH		
THEME	DEFINITENESS	DEF	
	ATTRIBUTES	HAS-COLOR	GREEN

Figure 2.3: Example of an event-based representation of Interlingua Model with the sentence *Mary did not slap the green witch.*

2.2 Automatic Speech Recognition

The mainstream approach to Automatic Speech Recognition (ASR) uses Hidden Markov Models (HMM) for pattern recognition over a large-vocabulary. Although HMMs will be described here in more detail there are also other approaches, such as maximum entropy models, segmental models, super-segmental models including hidden dynamic models and neural networks. In general, the approaches taken for pattern recognition of speech are based on statistical models. The reason for their popularity is that they allow the automatic parameterization from large training datasets, thus simplifying the development process and making it into a computationally feasible task [1].

Speech recognition have improved steadily its performance over the years, with a error-reduction rate of about 10% annually [22]. The state of the art solutions use large vocabularies with over 60 000 words, speaker independency, continuous speech and multilingual models. Research is done along three approaches, namely Knowledge based, Acoustic phonetic and Pattern recognition. The present work concentrates on the pattern recognition approach.

2.2.1 Architecture

Most ASR systems consist of a few basic components. See Figure 2.4 for an illustration, taken from [1]. The acoustic model includes the representation of acoustic, phonetics, gender, dialect and other features. The language model is used to calculate the probability of a sequence of words that are likely to co-occur. It tends to use the same models as in SMT, the n-gram models. The decoder undertakes the search of recognition hypotheses. The signal-processing module extracts salient feature vectors for the decoder, reducing the dimensionality of the input vector while maintaining the discriminating power of the signal. There is a number of techniques for the extraction of the vectors, such as: a) Principal Component analysis (PCA), which is a non linear feature extraction method; b) Linear Predictive, coding a static feature extraction method; or c) Cepstral Analysis, also a static feature extraction method used to represent spectral envelope [23].

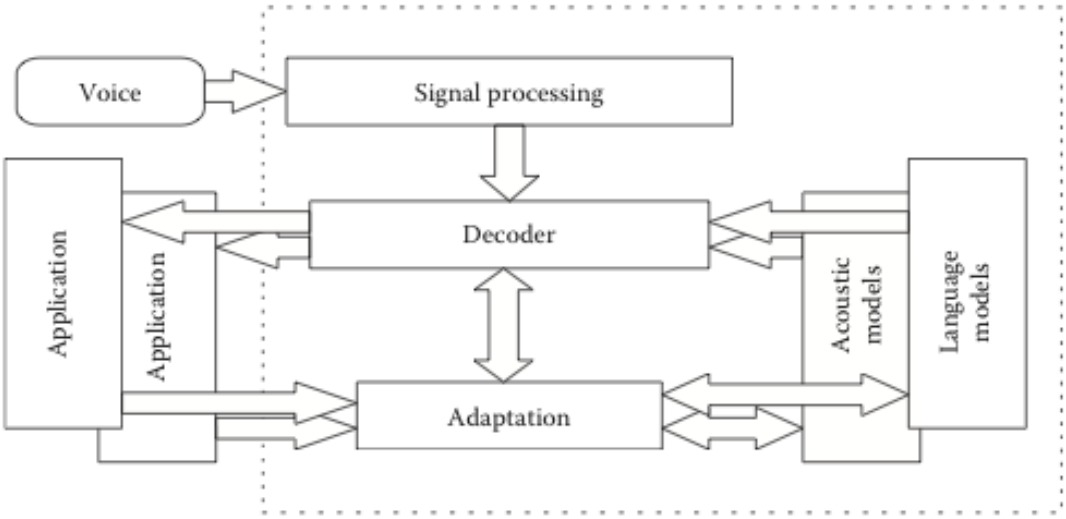

Figure 2.4: Typical ASR architecture (taken from [1])

2.2.2 Acoustic Models

The acoustic model plays a critical role in the accuracy of automatic speech recognition, by establishing statistical representations for the feature vector sequences computed from the speech waveform. The most influential factors over the accuracy are the context, speaker and environment variations.

In the context of S2ST, the acoustic model described here is of the class of continuous speech, thus allowing the recognition of continuous words.

The typical approach taken to create a acoustic model uses Hidden Markov Models.

A HMM is a finite state machine that allows to predict the likelihood of a hidden state sequence generating a state sequence. The transition between states is computed by a transition probability that stores the probability of entering a specific state given the previous state. In general, when using a HMM in speech recognition, each hidden state transition corresponds to a sound event, as a phone for example, that given an acoustic event, is to be infered. This inference is done by using an observation likelihood, that is, the probability of the acoustic event (observation) being generated from a state.

The challenge of this approach lays in the estimation of values for the state transition probabilities and the observation likelihoods at each state.

Typically, the Gaussian mixture model (GMM) classifiers are used to compute the observation likelihood, although there have been progresses with Deep Neural Networks, outperforming it in some cases [24].

Each entry in the acoustic models is created and trained on the basis of a lexicon containing the representation of each pronunciation along with the chosen basic phonetic units.

2.2.3 Language Models

Two types of language models can be used. A grammar or a probabilistic based language model.

The grammar based language model restrict the language to sequences of words that conform to grammar rules. It is typically used to recognize a small set of possible sentences.

The most used grammar based language models resorts to a context-free grammar (CFG). The CFG is a set of rules that recursively allows the definition of sequences of words. The probabilistic language model is used to calculate the probability of a given word sequence to occur as a sentence in a given language $[P(W)]$. This is usually done by resorting to n-gram models.

The n-gram model explores the frequency of a word occurring after a previous words. It is estimated using a training corpus with typically millions of words. It is created normally with sequences of just one word, a unigram $P(w_i)$, with sequences of two words, a bigram $P(w_i|w_{i-1})$, or more commonly with a sequence of three words, a trigram $P(w_i|w_{i-2}, w_{i-1})$.

Assuming the beginning and end of a sentence is marked with $< s >$, to calculate the probability of the example sentence *Goodbye blue monday*, $P(Goodbye\ blue\ monday)$, on the basis of bigrams, one takes

$$P(Goodbye\ blue\ monday) =$$
$$P(Goodbye| < s >)P(blue|Goodbye)P(monday|blue)P(< s > |monday)$$

The relevant frequency scores result from counting how many sequence of words occur. For example, for the trigram model, the probability of a word depends on the two preceding words. It can be estimated by observing the frequencies of the word pair $C(w_{i-2}, w_{i-1})$ and of the triplet $C(w_{i-2}, w_{i-1}, w_i)$, and then by calculating:

$$P(w_i|w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} \tag{2.4}$$

With a large vocabulary, an impractical number of n-grams may have to be considered. To address the problem of fine probability estimates to all of them, a smooth technique can be used, such as a linear interpolation of n-grams frequencies and a uniform distribu-

tion on the vocabulary, allowing to estimate the probability of all n-grams [25].

2.2.4 Decoding

The decoder is used to find the best sequence of letters and words that, in accordance to the acoustic and language models, correspond to the input feature vector sequence. This is done by searching the maximum posterior probability for the input feature vectors

$$\hat{W} = argmax_w p(W)p(X|W) \tag{2.5}$$

where the optimal sequence of words $\hat{W}$ given the sequence of acoustic inputs X is the sequence that, from each possible sentence, maximizes the product of the probabilities of the sequence, $p(W)$, according to the language model, by the probability of X given W, $p(X|W)$, according to the acoustic model.

The computation of the probabilities in the language model was already described in the previous section.

The main problem in the computation of the probabilities for the acoustic model relies on the vast search space, thus, requiring a search algorithm that computes only high probability states. The most common decoding algorithm is the time-synchronous beam search technique, which is a Viterbi decoder implemented with pruning. It performs a sub-optimal search by expanding the most probable paths at each HMM state and only keeping active those paths with likelihoods above a given threshold [26][1].

2.3 Text-to-Speech Synthesis

State of the art speech synthesis can nowadays achieve quite natural speech and is aiming for even more expressive and emotion associated speech. Although it is still distinguishable from a human speaker, the text-to-speech synthesis (TTS) is already far away from the ”mechanical” or ”robotic” sound it is most of the times still portrayed in popular media.

The main challenge of TTS is the conversion from a text representation to an acoustic waveform.

There are three main approaches in this area: Formant, Articulatory and Concatenative synthesis. As it is the mainstream approach, the concatenative method is the one described in this work.

In general terms, a TTS system using a concatenative approach chops up samples of speech fragments and stores them in a database, which are then combined and reconfigured to produce the sounds corresponding to sentences to be synthesized. It accomplishes

this task by, in a text analysis phase, first converting the input text into a phonemic internal representation, and in a second phase, called waveform synthesis, by converting that phonetic representation into a waveform.

2.3.1 Text Analysis

Processing the input text is done by a text analysis stage consisting of different procedures: text normalization, phonetic and prosodic analysis. The reason for analyzing the text lies in the dependency of the word pronunciation on the usage and the importance of the text for the modulation of the pitch, timing and amplitude of speech.

Text Normalization

The first procedure a TTS system uses is the normalization of raw text. The input text may come with abbreviations, numbers and acronymous that need to be normalized for their integration into the resulting speech.

The normalization begins by converting the text into tokenized sentences and finding the beginning and end of sentences. This is a seemingly trivial task but that carries a number of difficulties. As an example, not all occurrences of the punctuation symbol "." marks ends of sentences, they may mark also end of abbreviations.

Normally, the word tokenization is done with the application of a supervised machine learning method, being the logistic regression and decision trees the most common ones [6].

The second process consists in normalizing non-standard words. Words such as figures or abbreviations, need to be expanded so they can be pronounced. For example the number 2012 can be spoken in different ways depending on the context:

- Twenty twelve in "In the year 2012"
- Two zero one two in "The password is 2012"
- Two thousand and twelve in "It costs 2012 euros"

To undertake this task, the normalization of non-standard words procedure uses three steps:

- Tokenization to identify possible non-standard words;
- Classification to label words with their specific type (abbreviation, letter sequence, number, telephone, identifier, year and others; for a more complete list see [27]);
- Expansion of a token to a string of standard words.

For determining how to pronounce homograph words (e.g. desert and dessert, the first as a dry land and the latter as a after-dinner treat), the disambiguation is performed by running a Part-of-speech tagger to choose the pronunciation for a given homograph in its context. Then it is retrieved the distinct pronunciations of homographs from a stored database, labeled by Part-of-speech (lexical class). Some words may not be resolved by Part-of-speech taggers (e.g. the portuguese word *sede*, pronounced sÊde if meaning thirst, or sÉde if meaning headquarters), so alternatively they can be resolved by using word sense disambiguation algorithms like the decision-list algorithm of Yarowsky [28].

Phonetic Analysis

The phonetic analysis procedure undertakes the determination of the phonetic transcription from the normalized text. Most of the transcription is done using a large pronunciation dictionary. Dictionaries tend to offer just the word root pronunciation, not including morphological variations, like plural, feminine or conjugations.

The unknown words, mostly names, involves the use of morphological decomposition, analogical formation and mapping of unseen names to spelling variants.

Using the words orthography for the determination of the best sequence of phonemes is the most common approach. This approach, the grapheme-to-phoneme, converts a letter string like *DICTIONARY* into a phone string like *D IH K SH AH N EH R IY*.

The oldest technique involves the definition of rules written explicitly for the domain in use, typically context-sensitive rewrite rules. Most common nowadays is a data-driven approach using decision trees, pronunciation by analogy and neural-networks that computes the rules to be used. A third approach, with higher accuracy, is the statistical one [29].

Within the statistical approach, some techniques are used such as morphological parse trees using a layered bigram as statistical parsing [30], maximum entropy n-gram model and a joint maximum entropy n-gram model with syllabification [31]; joint-sequence models [32], Hidden Markov Models [29] or Conditional Random Fields (CRFs) [29][33].

Prosodic Analysis

Prosody is the study of the rhythm, stress and intonation aspects of a language. With it one can mark discourse structure or function, create saliency, and affective or emotional meaning.

The prosodic features create a segmentation of the speech chain into groups of syllables, called prosodic phrases. They can provide cues to signal the internal structure of the spoken message, as a slower or longer pause in a speech, or a change in the fundamental

frequency. To structure this knowledge, a lot from syntax, semantics and pragmatics is needed.

Although normally the speech synthesis takes a neutral approach to prosody, there have been research into sophisticated approaches that model the speech to obtain prosodic effects. Some of these approaches are the parameterization of intonation curves (F0) using an acoustic intonation model or acoustic stylization that express F0 curves as sequences of target points where transitions are filled by an interpolation function.

Two main trends are typical to take care of the prosody feature of the duration of speech. The most common assigns an average duration to each segment. The other makes use of large speech corpora and predicts the duration based on the observed data.

The result of all these steps in text normalization is an internal representation consisting of a phone string with prosodic boundaries and prominent syllables [6].

2.3.2 Waveform Synthesis

The main approach for Waveform Synthesis is the Unit Selection approach. The Unit Selection Synthesis [34] is a concatenative synthesis algorithm that generalizes from the diphone waveform synthesis method. Most modern TTS systems use it and it dominates the speech synthesis market with high quality and naturalness. The use of Statistical Parametric Speech Synthesis is a recent topic in terms of research [35].

The diphone waveform synthesis method consists of synthesis generated from a sequence of phones using prerecorded diphones. As the unit for concatenation, a diphone is an adjacent pair of phones representing the transition between phones [36].

It is impractical to use word-level concatenation given that a large amount of units would have to be recorded. It would be possible to record the syllables and phonemes of a language (for example, there are, approximately 10 000 syllables and 40 phonemes in the English language [5]) but due to coarticulation an unnatural speech is produced.

The synthesization process is done by selecting cut up diphones from a database where they were previously recorded from a person speech. The diphones are then concatenated and can be modeled for prosody (f0, duration) intentions.

The Unit selection method differs from the diphone waveform method by using *units*, pieces of speech that can be much larger than diphones, storing more than one record of each diphone and by not altering the prosody with signal processing techniques.

By storing a more extensive number of units in a database, the Unit Selection synthesis may obtain better accuracy considering that much of the words may already be present in the database thus resulting in a more natural waveform.

The units sequence selected from the database is done by taking into account the available unit sequences that meet the intended specification (in terms of F0, stress level

and phonetic neighbors) and that also performs the most smoothly concatenation without perceptible break.

The use of HMMs in Speech Synthesis have recently been a topic of research. The models can cope well with the large parameterization that within a Unit Selection system is limited in the variations of speech characteristics. The HMMs generate an approximation of similarly sounding speech segments, allowing flexibility in voice characteristics, speaking styles and emotions. Using techniques for adaptation such as interpolation eigenvoice (linear combination of vectors obtained offline from a set of reference speakers) and multiple regression [37] [38]. They are becoming very popular due to the high degree of flexibility that results from the statistical parametric representation of the voice.

2.4 Speech-to-Speech Translation

Speech-to-speech translation encompasses distinct fields of research. With the recent improvements of the distinct fields it is now possible to achieve S2ST with acceptable quality for some language pairs.

There is a vast number of S2ST systems. Some of the most notable projects are:

- VERBMOBIL, a now terminated long-term project aimed at the development of a mobile system for the translation of spontaneous speech in face-to-face situations.
- Global Autonomous Language Exploitation (GALE), a DARPA funded project whose goals were the development and application of computer software technologies to absorb, analyze, and interpret huge volumes of speech and text in multiple languages.
- Technology and Corpora for Speech to Speech Translation (TC-STAR), financed by the European Commission, was envisaged as a long-term effort to advance research in all core technologies for S2ST.
- Effective Multilingual Interaction in Mobile Environments (EMIME), financed by the European Commission, undertakes the development of a mobile device that performs personalized S2ST.
- IBM's Mastor, a S2ST system that translates spontaneous free-form speech in real-time on both laptop and hand-held PDAs. Developed with military focus for the U.S.A. allowing a bi-directional conversation with Iraquian people.

In S2ST the most common approach consists in just passing the output from one component to the next, where the components are of the type described in the sections above. The coupling of the components is done simply by the transmission of the single best hypothesis obtained from the previous component. Due to the peculiarities of spoken

language, it is proven that some other approaches can obtain a better accuracy. As an example of what can go wrong, with the wrong deletion of a word in the ASR process seeking to obtain a reduced word error rate (WER), the SMT result can be disastrous since the deleted word can be of key-importance to the correct translation. Transmitting more than one alternative from the ASR to the TTS has been an approach taken to improve the final result [39].

From Automatic Speech Recognition to Statistical Machine Translation

For a more accurate coupling, that is, for a reduction of the impact on SMT from possible ASR errors, techniques such as N-Best lists, Word Lattices or Confusion Networks can be used as the interface between the components. These techniques can improve the translation performance by transmitting more that one result from the ASR, passing along alternative results with their respective acoustic score [40].

As an example, the Word Lattice represents multiple word hypotheses, in a memory efficient way, where each word represents a link between nodes and have a specific time signature. It is also represented a start and terminal node, the begin and end of a sentence. For an illustration, see Figure 2.5, taken from [41] .

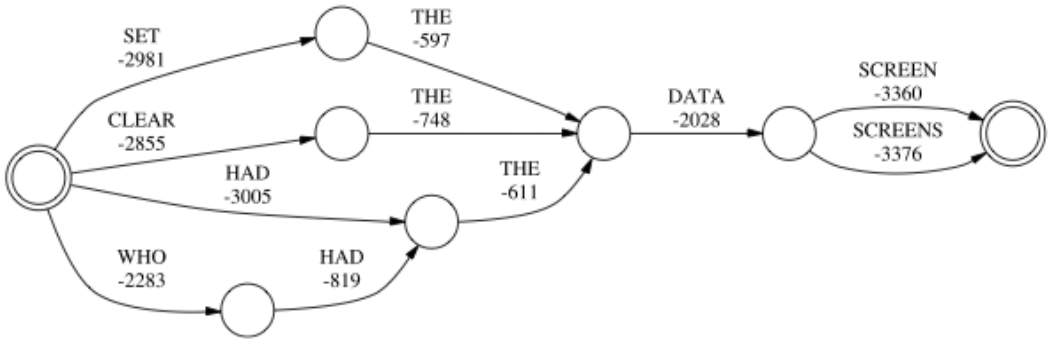

Figure 2.5: Word lattice example

There are also indications that the accuracy can be improved by incorporating the ASR acoustic and source language model scores into the selection process of the SMT system [39].

One other factor that can help the creation of a better coupling of the ASR component to the SMT component is the cleaning of spoken utterances. They form unnecessary words that do not contribute to the meaning of a sentence. Some techniques such as: noisy-channel models, HMMs, Maximum Entropy or Conditional fields have been used for the development of a cleaning component that removes these disfluencies [42][43].

From Statistical Machine Translation to Text-to-Speech Synthesis

In general, the coupling of the SMT component and the TTS component is done with the best translated sentence. More sophisticated approaches use the syntactic structure in the language generation process, and the dialogue context to improve the prosody prediction. This allows a better accuracy, naturalness and intelligibility since the TTS system depends greatly from the fluency of the translated sentence [40].

To handle possible wrong machine translations outcomes in the TTS system, some strategies can be applied with success [44]:

- pause insertion, where a pause is made so an untranslated word can be better identify and the sentence may become more intelligible;
- replacing untranslated words with fillers such as ”hum”;
- using the unit-selection-cost to identify bad joins and choose a better alternative from an N-best of the translation list to tackle bad phonetic joins.

2.5 Evaluation

Evaluating a S2ST system, or just one of its components is a controversial task but a necessary one. The main reason for the controversy is due to the subjectivity that is intrinsic to the tasks. There is no single perfect translation for every sentence nor a perfect speech pronunciations for all words. So the evaluation most of the times takes a qualitative approach in detriment of a quantitative one.

Typically a qualitative approach is done by human evaluators who conduct a survey over the results of a specific task, being asked to evaluate the adequacy, fluency, naturalness, translation or quality of speech.

Nevertheless, the quantitative approaches exist and are important to the iterative and rapid development of robust systems. They are specially promoted in the context of competition events. The National Institute of Standards (NIST) is one of the main promotors of such competitions in the USA that arrange for yearly competitions.

Automatic Speech Recognition

Evaluating an ASR system is typically done with the use of the Word Error Rate (WER) metric. A common measure is used that is derived from the Levenshtein distance or edit distance. The edit distance measures the quantity of insertions, deletions and substitutions necessary for the transformation of a string into another. The WER is the normalized result value of the edit distance from a reference word sequence and its transcription [45]. It can be formulated into:

$$\text{WER} = 100\frac{S + D + I}{N_r} \tag{2.6}$$

Where S, D and I are the number of substitutions, deletions and insertions respectively, and the N_r is the the total words in the reference word sequence. This means that for higher scores of these variables a worst performance result is obtained in terms of WER.

For example, given the reference word sequence *Goodbye blue Monday until other day* and the transcription *Goodbye you Monday until someday* there are a total of 6 words in the reference word sequence (N_r), 2 substitutions (*you* for *blue* and *someday* for *day*) and 1 deletion (*other*). This example results in the given WER:

$$\text{WER} = 100\frac{2 + 1}{6} = 50\% \tag{2.7}$$

Statistical Machine Translation

The simplest automatic methods for the evaluation of a translation are the precision and recall metrics. The precision metric computes the number of correct words of a translated sentence using a reference sentence.

$$\text{precision} = \frac{correct\ words}{translated\ sentence\ length} \tag{2.8}$$

The precision does not take into account the word order or the reference sentence length. For example, given the translated sentence *Goodbye blue Monday* and the reference sentence *blue Monday Goodbye maybe another day*, the precision would be perfect (precision=3/3).

In the recall metric the reference sentence length is used instead of the translated sentence length.

$$\text{recall} = \frac{correct\ words}{reference\ sentence\ length} \tag{2.9}$$

With the previous example sentences, the recall would be of 1/2, a 50% ratio. These metrics are commonly used together, using precision to output a ratio of the most correct words and the recall to check if some words maybe missing from the translation.

Like the ASR, the SMT can use the WER for the evaluation but since it does not take into account possible synonimous words and the reordering, other methods are preferred.

The most common evaluation technique is the Bilingual Evaluation Understudy (BLEU) metric. Like the precision and recall metrics, this metric compares the translated sentence with a reference sentence but instead of using words for comparison it uses n-grams obtained from the sentences.

With the matching n-grams, the $precision$ is computed in relation to the total number of generated n-grams of the same order i. These generated n-grams are computed through different n-grams orders and typically up to the fourth order. Different weights (λ_i) can be applied to the different n-grams order. Though typically they are set to 1. This model is the so called n-gram precision. The BLEU metric also implements a penalty for deleting words, the brevity penalty, since it reduces the score when the output is too short.

The BLEU metric for a n-gram with the maximum n order is defined as

$$\text{BLEU-}n = \text{brevity penalty} \exp \sum_{i=1}^{n} \lambda_i log\ precision_i \tag{2.10}$$

$$\text{brevity penalty} = \min(1, \frac{\text{translated sentence length}}{\text{reference sentence length}}) \tag{2.11}$$

An alternative method is the Metric for Evaluation of Translation with Explicit Ordering (METEOR). This metric was developed on the basis of the BLEU score and outperforms it in the measure of adequacy and fluency. The METEOR metric puts a strong emphasis on recall, taking into account synonyms or semantic closed related words. The downside of this metric is the complexity and effort it requires to create and evaluate a data set.

The main evaluation campaigns for SMT systems are the NIST, IWSLT, TC-STAR and ACL WMT. All of them tend to focus on a specific discourse domain.

Text-to-Speech Synthesis

The evaluation of a TTS systems is normally done by a qualitative approach, playing a speech synthesization and asking a group of listeners to rate the naturalness and intelligibility of the speech.

The Blizzard Challenge is a recognized standard TTS testing for the TTS research community. The TTS systems are tested with the same source material and in the same

time. Recently the scores are obtained from listeners using the Amazon Mechanical Turk platform [46].

2.6 Speech-to-speech translation in the Medical Domain

A number of Speech-to-Speech translation systems have been researched with a focus on the medical domain. The medical domain in speech translation have also been object of dedicated workshops at conferences [47].

Most of the S2ST systems use a bi-directional language translation between English and another language. The systems tend to be multimodal and to undertake the interaction as a doctor-patient dialog. In some solutions, it is only doctor oriented, allowing just the physician to pose questions, expecting a yes or no answer from the patient [48].

The S2ST systems are in general developed for a computer platform but recently there have been systems designed specifically for mobile devices since the portability have become one of the major convenience for its operation.

The goal of accuracy and broad coverage is always present in the conception of a S2ST system, but the fact is that for a better accuracy the domain needs to be severely restricted. Nevertheless, even using a strict domain, the medic context carries an expected critical accuracy regarding the safety of the translation and the information it conveys for the health of the patient. This critical accuracy implies an even more rigorous and concise domain definition.

It is known that the machine translation component is the weakest link among the three components, ASR, MT and TTS [49]. The lack of large enough language datasets for speech processing and translation in the medical domain is pointed as one of the main obstacles to its development. The scarcity of data affects specially the statistical machine translation component, whose strengths are already more linked with robustness than precision [50].

These are some examples of state of the art S2ST systems using a medical domain:

- IBM MASTOR [51] is a laptop and hand-held PDA system designed for medical emergencies. Developed for DARPA CAST and the Transtac programs with the mission of enabling rapid deployment of real-time S2ST for low-resourced languages. It supports regional dialects of Arabic and English and also Mandarin Chinese to English in a bi-directional way.

  In each of the components, a great attention to memory footprint is taken, taking into account that the mobile devices are more limited in terms of computational processing. The ASR component uses grapheme based acoustic models and it is based on the IBM ViaVoice engine that uses rank based acoustic scores. The MT

component is done with statistical phrase-based translation using weighted finite-state transducers for memory efficiency and speed. The TTS component uses a phrase-splicing and variable substitution method applied with embedded concatenative technology.

The training data and language model data is based on 200k utterances for the Arabic language and from web sources for the Mandarin Chinese.

- TRANSONICS [52] is a bi-directional dialogue translation system for the English and Persian languages. The ASR component couple with the Machine Translation (MT) component through a n-best lists/lattices with the decoding confidence score. The MT is a hybrid component with a table-lookup for approximately 1400 classes of sentences and a SMT. Using a doctor-oriented interface, after the speech is recognized and sent to the MT, the MT presents to the user the translation hypotheses along with the confidence and paired class.

  For the ASR components, different models are used. The most successful one happens to be the sub-phonetic mapping of English phonetics and Persian. The MT component starts by classifying the source sentence and by using a lookup table for translation, and when a certain confidence threshold is achieved with the classification, an already translated sentence is returned. If not, the SMT is used. The TTS use a hybrid Unit selection. When the MT translation results from the lookup table, the generated sentences are known a priori. When the translation results from the SMT, the speech synthesization is generated using a unit selection through diphone concatenation.

  The data collected for the necessary data sources, language models and the lookup table, was acquired from multiple sources, including medical phrase books. The most relevant data source was developped in collaboration with the USC Keck School of Medicine, with the Standardized Patient data collection translated to Persian.

- MedSLT [53] is a doctor-patient diagnosis dialogue system that focus on the use of grammar-based medical speech translation. It uses as subdomain the knowledge of headache, chest and abdominal pain. It supports any-to-any translation of Japanese, English, French and Arabic. It also supports a bi-direction English to Spanish version.

  The rule-based machine translation system used a data-driven corpus comprising recorded and transcribed utterances for the definition of its rules. A statistical machine translation is used in recent versions of the MedSLT, serving as a second recognizer, helping to solve out-of-coverage utterances. The ASR uses a Nuance Toolkit and all grammars are compiled with the Open Source Regulus platform.

The MedSLT uses around 500-1000 sentences per language and subdomain.

Chapter 3

Technology

I conducted a technology survey in the search for the most suitable set of tools that could support the construction of a bi-directional Portuguese-English S2ST prototype for the medical domain, focusing on the three main components (ASR, SMT and TTS). The necessary technology for the integration of the system along with the web and mobile application development was also considered. In the present chapter, I report on these surveys.

Nowadays, a range of speech technology software is available, from commercial solutions to free and open source software. New emerging solutions are available as web services, through API interfaces. Some of these solutions are even free although very limited in terms of the facilities they may offer for customization and personalization.

Since there can be quite a number of tools that can be listed, criteria were followed for their selection, basing it upon the state of update of the software, upon its level use by the research community or industry, and upon its adaptation to the intended technological requirements of this work. Not all tools have specifications available or that could be found. Typically, the evaluation score of the tools or systems is the specification that lacks in almost cases where specifications are incomplete, or totally missing. The Appendix E has a more detailed overview of the specifications for the addressed tools.

In the following text of this chapter, I will cover the survey for each of the necessary components to the construction of a S2ST prototype and the solutions eventually opted for.

3.1 Automatic Speech Recognition

Most of the solutions surveyed tend to use HMMs and apply it to support a continuous speech recognition system. The first criterium for the selection of the system to be adopted was the choice of a free solution. This decision left out two major solutions for ASR systems, which are the main commercial solutions that dominate the market:

- The NUANCE is the market leader in the industry, with a diverse set of tools for specific speech recognition environments and applications. Over the last years, NUANCE acquired dozens of other speech recognition companies, such as SVOX, IBM ViaVoice, Philips Speech Recognition Systems, Loquendo and SpeechWorks. It offers solutions for personal use, business, healthcare and developers. The most common solution, the Dragon Naturally Speaking Professional, offers speech recognition for various languages, including the Portuguese and English language. Its price is 480 Euros. The Healthcare solutions offer Physician speech recognition solutions, capturing the physician narrative, including medical vocabularies covering nearly 80 medical specialties and subspecialties. NUANCE claims a 99% accuracy out-of-the-box.

- The LumenVox is one of the market leaders in the industry and the second larger speech recognition solutions provider. It derived from the CMU Sphinx. It is a still active solution that is used in the market of small to mid sized businesses with solutions implemented in telephone systems, banking applications, call centers, outbound dialing and teleconferencing. The price plan for the ASR solution is of 2,800 Euros, offering the licenses to use along with technical support and training. It uses SRGS for the language modeling. No acoustic model exists for the Portuguese language. It also offers a software development kit (SDK) solution.

The second criterion for selection took into account the availability of an English and Portuguese acoustic model for the speech recognition engine. That criterion led me to left behind most configurable systems, mainly used in academic research. They are listed below:

- The Hidden Markov Toolkit is an open source ASR acquired by Microsoft and licensed to Cambridge University Engineering Department. No products can be sell based on HTK. It is available only for research purposes. It allows various language models and customization but no acoustic models are provided. It seems to be a on hold project, not delivering any new public update since 2009.

- Julius is an open source speech recognition system developed at the Nagoya Institute of Technology, in Japan, since 1997. It is a still active solution mainly used for academic research or as an ASR application for individual Unix users. It is highly configurable and supports various language models, from grammars to n-grams. It is also a highly portable system, being entirely written in the C programming language.

- The CMU Sphinx is a speech recognition system developed at Carnegie Mellon University. It offers an open source system with one of the largest and more complete set of tools for the creation of an ASR system. It is supported by a very active

community since 2000, being used mainly for academic research, but that spawn also to some commercial solutions. It also provides a thorough documentation and good community support.

Two systems conform to all the previous criteria:

- Google Voice is an API for the Google speech recognition system. It is only supported using the Google Chrome web browser and using the W3C Web Speech API Specification. It supports the Portuguese and English languages. Almost no configuration can be done with the system and there are no solid documentation or product specification.
- The Microsoft Speech Platform (MSP) is a set of speech tools that Microsoft makes freely available to developers with a SDK. It supports both the recognition of English and Portuguese languages. The system engine is only customizable in the language model component, allowing the use of a Speech Recognition Grammar Specification (SRGS) or a Statistical Language Model (SLM). The MSP provides complete documentation covering the development of an ASR system and can easily be integrated with web services.

3.2 Machine Translation

All solutions that required some payment were left out. Most of such systems offer excellent translation systems and are used in business solutions. In the list just below these systems are indicated:

- SYSTRAN is a commercial solution for a hybrid, rule-based and SMT, system. It can cost from 150 Euros to 11,500 Euros. The company SYSTRAN is one of the pioneers in the development of commercial machine translation software. Yahoo uses SYSTRAN for its Babel Fish translation. The company is highly focused on a few customers but has general public solutions with solutions from 70 Euros, like the SYSTRAN 7 Home Translator, to the 900 Euros, like the SYSTRAN 7 Premium Translator that supports English and Portuguese translations.
- PROMT is a spin off from the Herzen University, in Russia. It is a commercial solution with a strong Russian language translation component although it also provides English and Portuguese translations. It uses a rule-based machine translation system or a hybrid system with SMT. It offers solutions from 39 Euros, with the PROMT Personal 9.5, to the PROMT Professional, worth 1,200 Euros. It provides a line of machine translation solutions for key industries in Russia, as the sectors of Oil and Gas, Metallurgy or IT and Telecommunications.

- The SDL Language Weaver is a commercial solution for SMT. The system supports the translation of 24 bi-directional language pairs, including English and Portuguese. It was created using a phrase-based approach. For business integration, it offers customization for existing pre-translated data from clients. Several business oriented solutions are offered, from Aerospace to Government. The price range of the solutions range from 2,500 Euros with the SDL Tradys Studio 2011 Professional, targeting translators, to the SDL Trados Studio 2011 Starter, worth 100 Euros.

- Google Translate is a limited free but closed SMT service. For an extensive use or API integration, it is necessary to acquire a paid solution. It supports translation both Portuguese and English language. The information over the system is not publicly known. It offers alternative solutions to translations and it can automatically detect the input language. Google is enabling its translation services by offering, translation solutions for the millions of youtube videos, through collaboration with companies like Gendo and Translated.net.

- Bing is the translation service of Microsoft. It is a commercial application that offers limited free translations. It uses a statistical approach. The system translates between Portuguese and English languages. It offers an easy integration with other Microsoft products, like Office and Internet Explorer. It target normal users, developers and translators. The developers can use a web service for the translation request, where the prices can be from 30 Euros for a total of 4,000,000 characters/month, to 4,500 Euros for a total of 1,000,000,000 characters/month. No configuration can be done on the system.

- The WebSphere Translation Server is the IBM automatic translation system solution. It offers, from 20,000 Euros, an integration solution for a web infrastructure providing content translation. It supports the Portuguese and English languages. Due to IBM prolific use in the industry, this solution is mainly used by companies that already have a WebSphere Solution.

The next selection criterion was to opt for open-source systems that employ a statistical approach and that would provide the necessary documentation with the support of a well stable community. This criterion left out the Apertium system:

- Apertium is an open source rule-based machine translation. Funded by the Spanish government for the OpenTrad project. Currently being developed by Transducens, from the Departament de Lienguatges i Sistems Informàtics of the Universitat d'Alacant. It provides an easy creation of machine translation systems with

the use of a well-specified XML format. The translation for Portuguese and English languages is provided. The project is still active with recent fundings from the Spanish government.

Two systems came out from the selection. Both are open-source and employ a statistical approach, with a focus on the academic research:

- Moses is an open source statistical machine translation system. It is mainly supported by the EuroMatrix project, funded by the European Commission. It is a very active project, being largely used by research groups. It implements state of the art statistical approaches, as phrase-based and tree-based models. It can be trained for any pair of languages provided there exists the necessary data. It makes it possible to implement a word-lattice for the coupling with an ASR system. The project have a strong community and the tools a vast documentation.

- NiuTrans is an open source statistical machine translation system developed at the Northeastern University, in China. It supports phrase-based, hierarchical phrase-based and syntax-based models for research-oriented studies. The system is relatively new and active. Its first version was made available in 2011, and the latest, still in beta, is available since January 2013. The system allows the configuration for the languages I may want to train it with.

3.3 Text-to-Speech Synthesis

Likewise the previous components, the first criterion for the selection of a TTS system was that it offer a free solution. This criterion left behind the following TTS systems:

- The Loquendo Text to Speech is the NUANCE speech synthesis system. It is available for over 30 languages, with more than 70 voices across major operating systems. It allows the creation of user Lexicons, prosody controls and supports the Speech Synthesis Markup Language (SSML), a markup language for assisting the generation of synthetic speech. Both Portuguese and English languages are supported. This solution is employed in a set of different NUANCE products, like the PDF Reader or Nina, a virtual assistant for customer service.

- Cepstral is a company that provides speech synthesis technology since 2000. It was a spin off from Carnegie Mellon University and nowadays provides speech synthesis from mobile devices to centralized servers. The standard price is about 155 Euros for a individual use, with a 30 Euro plan for each voice. No option for the Portuguese language is provided. One of the main focus of the company has been the production of voices, most of them being used by other speech synthesis companies.

- The IVONA is a speech synthesis system recently acquired by Amazon for its kindle product range. It uses a concatenative approach with Unit Selection. In the Blizzard Challenge of 2006, it won with the highest score. It supports English and Brazilian Portuguese. The solutions it provides include cloud services and web services for individual use as computer reading for documents, news, rss, books and emails or even film voice-over. There are also business solutions, offering a SDK, a speech cloud, a speech Server or a Telecom solution.

- Acapela is a commercial speech synthesizer that result from joint efforts of three speech companies: Babel Technologies, Infovox and Elan Speech. It comprises both Portuguese and English languages, along with over 100 voices in 30 languages. Its focus is on the easy creation of voices for a wide range of applications, from accessibility and assistive solutions for Dyslexia and cognitive impairments to screen readers, audio publishing (books, content newspaper), consumer electronics (robots, toys), education, industry, etc.

- The Natural Voices is an AT&T commercial speech synthesizer. The AT&T research group works in the TTS field since 1998 and since then it gained some public highlights, being used in movies like Red Planet and I Am Sam. It does not cover the Portuguese language, but includes the English language. It is used in customer support dialog systems, interactive voice response, unified messaging and screen reader programs. The project is still active with recent research publications. A solution can cost up to 800 Euros.

Naturalness of speech was the second criterion of inclusion, so a concatenative approach was needed. Two systems were thus left out:

- The eSPeak is an open source formant synthesis speech synthesizer. It is written in C and available for a large set of operating systems. Open source development is mainly done by Jonathan Duddington. The system is characterized as being fast but with a very artificial pronunciation. It allows both English and Portuguese languages. Since it has a small size, it is mainly used in mobile devices or machines with weak processing power.

- Praat provides software that permits an analysis of speech in terms of it phonetic representation, and that also allows speech synthesis using formant or articulatory synthesis. It is still an active project from the University of Amsterdam. Only the English language is supported and its main use is for academic research. It can be implemented across platforms, being written in C and C++.

The need of a component that could cope with the synthesis of the Portuguese language was the third criterion. Accordingly, two open source systems were left out:

- The Festival speech synthesis is an open source system developed at the University of Edinburgh. The last release was on 2010. The English language is supported but the Portuguese is not. It uses a concatenative synthesis and allows the creation of different language models. It allows a high degree of customization and is mainly used for academic research.

- The MARY is a Java open source speech synthesis developed at Saarland University. It uses concatenative synthesis with HMM-based synthesis voices. MARY is provided with a large set of tools with different functionalities, from part-of-speech tagging to intonation annotation. The system is mainly used for academic research. It supports the English language but not the Portuguese language.

Only one solution was found that complies with all the selection criteria:

- The Microsoft Speech Platform provides a free to use text-to-speech synthesis. Both English and Portuguese languages are supported. The system uses a concatenative approach. It does not allow a high degree of customization, except for the possibility of generating speech from SSML. The MSP provides a complete documentation on the development of a TTS system that can be easily integrated with web services.

3.4 The options chosen

Taking into consideration the technological solutions available and the objectives of this work, two ASR systems were chosen, the Microsoft Speech Platform (MSP), for the mobile application, and the Google Chrome ASR, for the web application. The MSP was also chosen for the TTS system. For the SMT system, Moses was chosen.

The choice of Moses for the SMT system was done taking into account the large amount of research papers focused on the system, the fact that it is a state of the art SMT system allowing coupling, with web services integration via shell interface and the fact that is highly customizable.

The main factor for the selection of MSP as the ASR system is the already available packs for the Portuguese and English languages along with the easy creation of and integration with web services, the possibility of creating a SGRS compliant grammar, the heavy documentation and the possible coupling with SMT.

The choice of MSP for the TTS system was favored by the free license, the concatenative approach, the availability for the Portuguese and English languages and also given the rapid integration with web services it permits.

Although the systems implemented with the MSP integrated non open software, the use and deploy of an application with it can be made free, since the API is a freely-redistributable component according to its license.

For the web application, the Google solution was chosen since it is the only browser that implements the W3C Web Speech API Specification for speech recognition. It does not allow for any customization except the choice of language to be recognized. The web application needed to be implemented specifically for the Google Chrome browser, as no other browser natively allowed speech recognition by the time this survey was done.

All the other solutions fell short in terms of language adaptation blocked, poor customization facilities or practical unavailability because of the associated license or costs.

The decision concerning the framework for the mobile application fell on the iPhone, since my research unit (NLX-Group) had one such device available for experimentation and development.

Since the Moses system needs a UNIX platform to run and the MSP a Windows one, the communication was to be done over web services, a method of communication between two encapsulated programs over the web.

Chapter 4

Implementation

In this chapter I will start by providing a general overview of the implemented S2ST system. Next, I will describe the EUMed Corpora, the corpora used for the creation of language models and the training of the SMT models. I proceed with the description of the implementation for each component of the S2ST system, the integration of the components through the use of web services, and finally the description of the implementation of the web and mobile applications.

An overall perspective of the S2ST mobile application can be found in Figure 4.1. When an English or a Portuguese user speaks to the phone, the other person will subsequently listen to the translated speech.

Upon a user request, the application starts the speech recognition by recording the speech and sending it as a stream of audio to the ASR web service. In response, the ASR web service returns the speech recognition result as the sentence with the most confident score.

With the resulting sentence, the mobile application calls the translation web service and in response gets the respective textual outcome, in the target language of the SMT component.

With this raw text outcome, a speech synthesization is requested to the TTS web service. The TTS web service returns a stream of audio representing the speech in the target language. The audio stream is then played in the mobile phone.

When using the web application, in turn, the main difference occurs with the use of a different ASR system, as can be seen in Figure 4.2. In this case, the recording and speech recognition uses the Google Chrome implementation of the W3C Web Speech API Specification.

With the resulting speech recognition the web application requests the translation into the target language and then the speech synthesization, receiving the former in raw text and the latter in audio format, as the mobile application does.

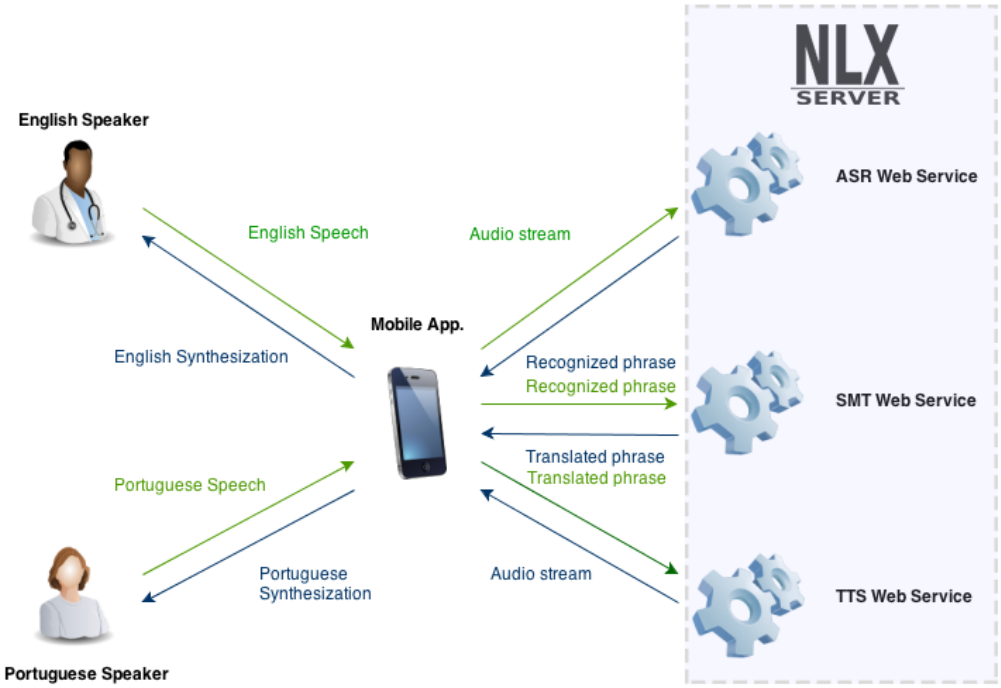

Figure 4.1: Overall perspective of the S2ST Mobile Application

Figure 4.2: Overall perspective of the S2ST Web Application

4.1 EUMed Corpora

A first step towards the implementation of the components was the gathering of parallel corpora in the Portuguese and English languages. The choice for these languages was taken given the existing volume of data available for both languages (larger than for any other pairs involving Portuguese) and also my knowledge in reading and writing of these languages.

This is one of the most important steps for the construction of an efficient S2ST system, especially when using statistical models. This is due to the need to use the parallel corpora for creating language models and training the SMT system.

Since a larger volume of data improves the statistical approaches, a large collection of parallel texts was convenient.

Taking also into account the domain, I searched for corpora in the medical domain and especially of doctor-patient dialogues. However, there exist very few available parallel corpora in the domain of medical diagnosis and none consisting of doctor-patient dialogues.

Alternative data were thus needed to obtain fluency, independently of the domain.

In the next list, I present a list of the corpora used and their specification:

- The Europarl [18] consists of a parallel corpus extracted from the proceedings of the European Parliament. It includes 21 European languages in which the Portuguese and the English languages are included. The corpus was already sentence aligned and consists of 1,960,407 sentences for each language.

- The JRC-Acquis [19] is a body of European Union law applicable in the EU Member states. It covers 22 languages, in which Portuguese and English are included. This corpus needed to be converted from its original format and sentence aligned. The result was 1,236,846 Portuguese-English aligned sentences for each language.

- The Unified Medical Language System (UMLS) medical Metathesaurus is a large, multi-purpose, and multi-lingual thesaurus that contains biomedical and health related concepts. It uses Mesh, RxNorm, LOINC and SNOMED CT medical databases as sources. I requested from the U.S. National Library of Medicine a license to use the UMLS [54]. After installing the UMLS, I extracted and aligned the Portuguese and English database entries. The resulting corpus consisted of 85,501 aligned sentences for each language.

- The EMEA is a parallel corpus made out of PDF documents from the European Medicine Agency [55]. It contains description of medication usage, treatments and diseases. The resulting Portuguese-English aligned corpus consisted of 1,082,144 sentences for each language.

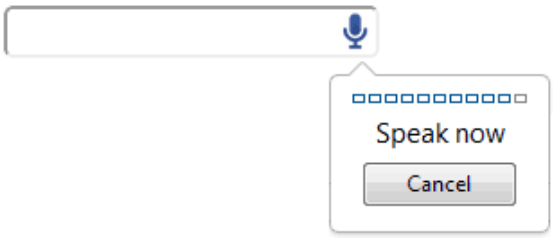

Figure 4.3: Snapshot of the web application speech recognition UI

- The ECB corpus consists of European Central Bank documentation [55]. The Portuguese-English parallel corpora used contained a total of 202,030 aligned sentences for each language.

4.2 Automatic Speech Recognition

Two ASR components were used, the ASR component for the web application and the ASR component for the mobile application.

The ASR system for the web application used the W3C Speech Input API Specification implemented in the Google Chrome. To proceed with the speech recognition, I implemented the specific W3C Speech API tags through HTML programming. These tags generate in the user interface (UI) a button that allows to initiate a sound recording event, as depicted in Figure 4.3. The main tag is represented as <input>, with these specific attributes:

- `x-webkit-speech`, creates an instance of speech recognition.
- `lang`, defines the language for the speech recognition, this attribute is dynamically changed with a Javascript to accommodate the desired language for speech recognition.
- `onwebkitspeechchange`, defines the specific event that will handle the resulting speech recognition output. The event routine was implemented in Javascript.

The user may request a speech recognition by pressing the UI microphone button and speaking to a microphone. The browser implements the speech recognition by using the Google Speech API.

The remaining of the present section cover the ASR system used by the mobile application.

The ASR system used by the mobile application was implemented using the Microsoft Speech Platform tools (MSP). The major development effort took place with the MSP Server tools (speech recognition and speech synthesis), along with the Runtime Languages for American English and European Portuguese. These language packages provide the acoustic models. Most of the programming was done with the C# language using the MS Visual Studio 2010.

I contacted Microsoft, through its Language Development Group (MLDC) in Lisbon, to obtain a deeper knowledge of its ASR components. A general description of the acoustic model was provided as follows: " The acoustic models are HMM-based. They comprise a mix of gender-dependent whole-word models and cross-word triphones trained using several hundred hours of read and spontaneous speech collected from adult speakers of European Portuguese and American English. It uses a silence model, a hesitation model for modeling filled pauses, and a noise model for modeling human and non-human noises." (Personal communication).

For confidentiality reasons, Microsoft did not publish any papers on this matter and it is not possible to change the current acoustic model or the decoder of the ASR implementation. Although the decoder and acoustic model are closed for configuration or analysis, the language model component can be accessed.

Language Model

The MSP ASR component accepts as a language model the Speech Recognition Grammar Specification (SRGS) or a Statistical Language Model (SLM). The SRGS is a W3C standard that allows the definition of speech recognition grammars. It has the expressive power of a Context Free Grammar (CFG).

Only the defined words in the language model and the underlying structure between them were used by the ASR system to form a possible candidate of the speech recognition.

As a exploratory experience both language models were used for an initial assessment to define the best approach. Three experiments were made. In the first experiment, I implemented a rudimentary grammar model. In the second experiment, I implemented a more complex grammar model. And in the third experiment, I created the language model with a SLM.

The experiments will now be described and will be followed with assessment remarks.

The **first experiment** was done with a simple grammar model. The grammar was implemented with the SRGS, using an XML format.

To support the design of this grammar, I created a hypothetical doctor-patient dialogue (Anamnesis Corpus, that can be seen in Appendix A), in Portuguese, to address the need of a specific doctor-patient diagnosis and treatment corpora. The dialogue is based on

a typical case of a minor car accident resulting in an arm injury. It takes into account the general procedures when handling this type of cases, comprising the physician and patient responses. The assembled dialogue had 160 sentences and the methodology and typical dialogue context were done with the help of an experienced family physician. The physician helped me by applying the anamnesis (medical history) process for the production of the medical diagnosis.

The anamnesis is the process used by physicians to gather the necessary information to formulate a diagnosis and provide the medical care necessary to the patient. It consists in asking key questions to the patient allowing the inference of the patient symptoms.

Based on the physician experience and past cases, the typical answers of the patient were added along with the ramifications of the physician questions to achieve the necessary medical diagnosis of our case, the arm injury resulting from a car accident.

The following text is an example of a possible dialogue:

```
Physician: Good Morning, what brings you here?
Patient: I have a sore arm, I had a car accident.
Physician: In which arm?
Patient: Left arm.
Physician: What part of the arm?
Patient: Elbow.
Physician: Does the pain radiates to somewhere else?
Patient: No.
Physician: Can you move your arm?
...
```

The dialogue was structured in different phases so the ramifications of possible questions and answers could be included according to the anamnesis process:

- Greetings and Presentation: utterances of introduction between two persons. Examples: *Good morning*, *Goodbye*, *My name is ...*, etc.

- Introduction: utterances that present the problem of the patient. Examples: *What brings you here*, *My arm hurts*, *I had an car accident*, etc.

- Assessment: utterances that allow the physician to assess the patient situation. Examples: *Does this hurt*, *Try to bend your elbow*, *Can you move your fingers*, etc.

- Investigation: exposition of the situation and necessary continuation of observation with other medical process. Examples: *You will need a X-Ray*, *You may have a broken arm*, etc.

- Planning: informing the patient of the situation. Examples: *You have nothing broken*, *I will prescribe an analgesic*,etc.

- Common utterances: general utterances employed in conversation. Examples: *No*, *Yes*, *Maybe*, *Please*, etc.

A set of rules were defined to form a rudimentary grammar that could cope with the Anamnesis Corpus.

The grammar was organized in sections to be in correspondence with the phases of the dialogue. Each phase was represented as a rule. In each rule a subset of items corresponded to the defined utterances.

Below, an example is offered representing the phase of assessment where two utterances are defined:

```
<rule id="assessment">
   <one-of>
      <item>
         <sapi:subset sapi:match="subsequence-content-required">
           bend your elbow
         </sapi:subset>
     </item>
     <item>
        <sapi:subset sapi:match="subsequence-content-required">
           squeeze my fingers
         </sapi:subset>
     </item>
   <one-of>
</rule>
```

The utterances were tagged with the `sapi:match` attribute and the subsequence-content-required property, indicating that the recognition of a string, as for example the string *bend your elbow*, should take place when one or more words in the string are recognized as a sequence in the audio input.

The **second experiment** was done by using a more complete grammar construction process. It was implemented by first creating a CFG grammar (Med-CFG) and then by converting it to the SRGS language.

The CFG describes the structure of sentences with the use of production rules. These rules generate a set of sentences that consist in the language defined by the grammar. Along with the rules, there are also terminal and non-terminal symbols. The terminal symbols represent the elementary symbols of the language defined by the grammar. In the implemented grammar they consist of words. The non-terminal symbols represent expressions or phrases.

As an example of a CFG can be as follows:

```
Frase -> SN SV
Frase -> SV
Frase -> SN
SN -> Artigo Nome
SN -> Nome
SV -> Verbo SN
SV -> Verbo

Artigo -> [o,os,as,um,...]
Nome -> [braço,cotovelo,dor,...]
Verbo -> [dobre,coloque,imobilize,...]
```

In this grammar, all sentences defined through the `Frase` rule are accepted as part of the language. The grammar defines sentences made of a Noun Phrase and/or a Verb Phrase, where the Noun Phrase is composed of an Article followed by a Name, or just of a Name. The Verb Phrase, in turn, is composed of just a Verb or a Verb followed by a Noun Phrase.

An example of a partial conversion between the previous CFG rules $Frase$, SN and terminal symbol $Nome$ to a SRGS can be as follows:

```
<rule id="Frase">
 <item>
   <ruleref uri="#SN"/>
  </item>
  <item>
   <ruleref uri="#SV"/>
  </item>
</rule>

<rule id="SN">
  <item repeat="0-1">
    <ruleref uri="#nome">
  </item>
</rule>

<rule id="nome">
  <one-of>
    <item>braço</item>
    <item>cotovelo</item>
    <item>dor</item>
  </one-of>
</rule>
```

Although this grammar had a few rules, a large set of sentences could be generated from it. As an example, the sentence *imobilize o braço* would be part of the language generated by this grammar. See Figure 4.4 for its parse tree.

The Med-CFG and its conversion to SRGS, Med-SRGS, contains a more complete grammar than the example above. The complete grammars can be seen in Appendix B and Appendix C respectively.

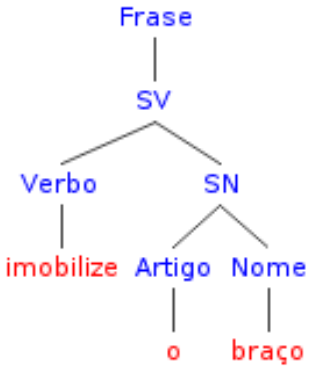

Figure 4.4: Syntax tree of the sentence ”imobilize o braço”

The conversion was done manually by rewriting the Med-CFG to the SRGS format. I encountered some limitations in the conversion of the CFG, being the most problematic the CFG left recursion, where the implementation solution in the Med-SRGS created a less rigorous syntax rule to accommodate the possibility of left recursion.

The **third experiment** was done using a Statistical Language Model (SLM). The SLM allows to assign a probability distribution to a sequence of words using n-grams. The MSP accepts a SLM in the ARPA-format, restricted to trigrams at most.

Two language models were created. A first one with the Anamnesis Corpus, with approximately 160 sentences, with a total of 445 words. And a second one with an extract of the parallel corpora described in the section of EUMed Corpora above. This extract had 30,000 sentences, with a total of 847.876 words. Both language models were created with a Portuguese corpus.

Both corpora were tokenized, lowercased, encoded and stripped of unspoken items such as question marks and citations. After cleaning the corpora, they were converted to trigrams using the SRI Language Modeling Toolkit (SRILM) that contains a executable program for the creation of n-grams.

The ARPA-format represents an n-gram entry with the words and their conditional probability (logarithmic with base 10). Some word sequences can also include a back-off weight associated. This is the probability that the following word for a n-gram is not defined by any of the (n+1)-grams.

As an example of an unigram entry, the word *onde* is represented as follows:
$-3.372356\ onde\ -0.8404543$

The first number represents the conditional probability and the latter number the back-

off weight. In the example seen above, the word *onde* would have the probability of $10^{-3.372356}$ of occurring and the assigned probability of $10^{-0.8404543}$ of not having a word following it in the bigrams.

As an example of a bigram entry, the phrase *no cotovelo* is represented as follows:
$-0.5037242\ no\ cotovelo\ 0.1317638$

In the example seen above, the word sequence *no cotovelo* would have the probability of $10^{-0.5037242}$ of occurring and the assigned probability of $10^{0.1317638}$ of not having a word following it in the trigrams.

All words from the selected corpus were added to the n-gram and the probabilities were estimated using the total counts. The back-off model used by the SRILM was the Katz back-off model.

As the **result of these exploratory experiments**, through a qualitative assessment, I decided for the SLM approach as the most successful one.

The first experiment, with the rudimentary grammar, allowed the speech recognition system to recognize the same spoken sentences with a high level of certainty. The downside is the strictness domain that it imposes and the subsequent rigidness. An out-of-corpus phrase structure would simply be ignored. To obtain good results with this approach, taking into account the doctor-patient dialogue dynamic, it would have been necessary to gather an much more larger corpus of possible sentences of an doctor-patient dialogue, and that is just unfeasible in the context of my work.

The second experiment, using a more complex grammar, allowed a more dynamic and larger scope of dialogue with the use of the same doctor-patient dialogue corpus. Nevertheless, it proved to be also an unattainable approach to yield good results. Although the realization of a more complete and embracing grammar could bring good results, the task is a very laborious one and also out of the scope of the present work. Another problem that emerged from this approach was also the higher time consumed by the MS ASR system during the speech recognition process.

The third experiment proved to lead the most robust results. When using the statistic language model, created solely with the doctor-patient dialogue, a very high level of certainty was found in the speech recognition. It also showed a much better performance.

The language model generated from the EUMed Corpora, in turn, was able to handle the doctor-patient dialogue sentences along with a larger range of sentences, while still obtaining a good level of recognition.

Due to simplicity and efficiency demonstrated by the SLM approach, it was chosen for the creation of the language model.

When experiencing with unigram, bigram and trigram, the latter yield better results. Since the ASR system needed to recognize two languages, Portuguese and English, for each language a specific language model was created from the EUMed Corpora.

After creating the n-grams, the resulting ARPA file was compiled to a binary file for optimization. This allowed to reduce the processing load in the ASR system and the application latency when loading.

4.3 Statistical Machine Translation

Moses, the chosen SMT system, is an open-source statistical machine translation system mainly supported by research funded by the European Commission. It comprises a collection of tools and utilities, whose main components are the training pipeline and the decoder. It supports different types of translation models: phrase-based, syntactic and factored. Both phrase-based and factored models were used in the creation of the SMT system in order to choose the most suitable one for my S2ST system.

The main steps to the creation of an SMT involves the preparation of the corpora, the creation of a language model, and the training and tuning of the resulting SMT component.

Preparation of Corpora

To train a statistical machine translation, parallel corpora needed to be created, cleaned and formatted. This revealed to be a very time consuming and less trivial than expected task.

Some corpora needed to be converted from its original format, as it was the case of the JRC-Aquis, where the original format was a XML, or needed to be extracted, like the UMLS. A special attention to the corpus is needed so that the necessary alignment of sentences is ensured.

In order to support the training of a factored SMT model, the corpora needed to include more linguistic information than the one used by the phrase-based model. The factored model takes into account the integration of linguistic information that could improve the success of the translation.

After the corpora were collected and sentence aligned, they needed to be annotated. Together with the surface form of the words, two linguistic dimensions were used to improve the translation.

The annotation consisted on the incorporation, in the parallel corpora, of the Lemma and Part-of-speech Tagging (POS) of each token. This annotation needs to be done for both the Portuguese and the English corpora. Since each language has different syntactic

```
<sentences>
  <sentence id="1">
    <tokens>
      <token id="1">
        <word>Mr</word>
        <lemma>Mr</lemma>
        <CharacterOffsetBegin>0</CharacterOffsetBegin>
        <CharacterOffsetEnd>2</CharacterOffsetEnd>
        <POS>NNP</POS>
      </token>
      <token id="2">
        <word>Prodi</word>
        <lemma>Prodi</lemma>
        <CharacterOffsetBegin>3</CharacterOffsetBegin>
        <CharacterOffsetEnd>8</CharacterOffsetEnd>
        <POS>NNP</POS>
      </token>
```

Figure 4.5: Excerpt of a text annotated with CoreNLP

and morphologic properties, different sets of tools were needed to perform this annotation.

The annotation of the English text was done by using the CoreNLP tools[56]. The process involved different stages:

- The tokenization: where the text was broken up into words and symbols as tokens.
- Sentence annotation: the resulting tokens from the tokenization stage were groupped into sentences.
- Part-of-speech annotation: each token was labeled with their Part-of-speech tag using the Penn Treebank tag set.
- Lemma annotation: for each token, it was associated with its Lemma.

The resulting annotated corpus was a structured XML file in which each sentence had branches for each token, and each token had branches for the word (surface form), the Lemma, and the Part-of-speech tag, as can be seen in the example in Figure 4.5.

The Penn TreeBank tag set has thirty six Part-of-speech tags. As an example, the word *translate* gets the tag VB, where VB means Verb in the base form; and the word *Portuguese* gets the tag JJ, where JJ means Adjective, in *The Portuguese Republic*, or the tag NN, where NN means Noun in the singular, in the context *He speaks Portuguese*.

For tagging the Portuguese corpus the LX-Suite tools [57] were used.
The process involved the following tools:

- The LX-Tokenizer: this tool segments text into lexically relevant tokens, using whitespace as the separator; expands contractions; marks spacing around punctuation or symbols and detaches clitic pronouns from the verb.

- The LX-Tagger: this tool assigns a single Part-of-speech tag, from the tagset, to every token.
- The LX-Lemmatizer and Featurizer: this tool assigns inflection feature values to words from the nominal categories, namely Gender (masculine or feminine), Number (singular or plural) and, when applicable, Person (1st, 2nd and 3rd); assigns degree feature values; assigns a Lemma to words from the nominal categories; assigns a Lemma and inflection feature values to verbs.

The resulting annotated corpus kept the original alignment with the English corpus. The corpus was annotated by adding, after each token, the Lemma (when the token differs from the Lemma) and the Part-of-speech tag using a slash as separator.

The LX-Suite uses more than fifty Part-of-speech tags. As an example, the word *traduzir* can get the tag V, where V means Verb. The word *portuguesa* can get the tag ADJ, indicating an Adjective in the following context *A República Portuguesa*, or the tag NP, indicating noun phrases, in *Ela é portuguesa*.

After both annotation tasks have been performed, each of the two parallel corpora, in Portuguese and in English, had to be converted to Moses input format. Two separated parallel-corpora were also created, one pair of corpora consisting only of Lemmas and the other pair consisting only of Part-of-speech tags. They were aligned along the alignment of the original surface form.

The next step was the filtering out of long sentences. Every sentence that had more than 80 characters was truncated. This was done due to a limitation in the word alignment tool. The alignment was done with word alignment toolkit GIZA++.

After the filtering was done, the corpus was then lowercased.

The resulting annotated corpora consisted of 3,906,126 aligned sentences for each language. The Portuguese corpus has 94,308,489 tokens, with an average of 24 tokens per sentence. The English corpus has 82,803,111 tokens, with an average of 21 tokens per sentence. Including the annotations (Lemma and Part-of-speech tags), the Portuguese corpora contained 343,188 word types and the English corpora 320,636 word types.

The preparation of the corpora just described was done with the EUMed corpora and used for the training of the SMT component.

For the tuning and evaluation I used the Official Europarl test set corpora made available in the ACL Third Workshop on Statistical Machine Translation [58].

These parallel corpora consist in recent proceedings of the Europarl, not included in the EUMed corpora. The evaluation corpora contains 2.000 sentences for each language

pair with 54.750 tokens for the Portuguese corpus and 54.302 tokens for the English corpus. And the tuning corpora contains 2.000 sentences for each language pair with 53.686 tokens for the Portuguese corpus and 53.536 for the English corpus. The reason for choosing this corpora was based upon the fact that it is the most recent parallel corpora in which the Portuguese and English language pair was evaluated, through the Euromatrix evaluation campaign, allowing for a comparison of results.

Language Model

The SMT language models follow the same principles that were used for the ASR component. For each language, models for the surface words and for the Part-of-speech tags were built. The surface word language models were used in the phrase-based model and in the factored model while the Part-of-speech tags language models were used only in the factored model.

The language models for the surface words used the original corpora, resulting in a file containing the unigrams, bigrams and trigrams. The following text is an excerpt of one of the English language models created using the surface words:

```
\data\
ngram 1=424464
ngram 2=5122665
ngram 3=6731815

\1-grams:
...
-4.650786 injured -0.5790805
-5.898272 injures -0.1440469
-5.548884 injuring -0.1657944
-4.814766 injurious -0.544678
-5.835898 injuriously -0.5109516
-4.173601 injury -0.6755068
...

\2-grams:
...
-4.036222 <s> injury -1.390034
-2.924922 abdominal injury -0.2335182
-1.500642 accidental injury -0.7921941
-3.429111 an injury -0.9505887
-4.227573 and injury -0.8965784
-2.200521 ankle injury -0.2441482
...
\3-grams:
...
-2.000387 <s> abdominal injury
-0.949074 injury accidents by
-1.426936 bleeding after injury
-2.702826 after an injury
```

```
-2.039812 have an injury
-1.424693 suffered an injury
...
```

The language models for the Part-of-speech used the file consisting only of Part-of-speech tags. The following text is an excerpt of one of the English language models created using the Part-of-speech tags:

```
\data\
ngram 1=47
ngram 2=1924
ngram 3=37047
...
\1-grams:
...
-0.8354089 nn 0
-1.490095 vb 0
-2.614345 pos -2.012941
...
\2-grams:
...
-0.7982951 cc nn 0
-1.471827 cc vb 0.829314
-1.674174 nn cd 0.3785322
...
\3-grams:
...
-3.698358 nn pos </s>
-0.5900444 vb pos vb
-1.691476 vbn wdt dt
...
```

For the creation of the language models, the SRILM tool was used. The language models were based on trigrams, using interpolation with the Kneser-Ney discounting.

Train

With the parallel corpora and language models in place, the training process could be performed. This process was demanding in terms of time and computational resources, and may take at least a week to be completed when no experimenting is involved. Each directional pair of languages, English to Portuguese and Portuguese to English, needed to be trained, resulting in the construction of two decoders.

The training of the SMT component was done using the Moses tools, except for the GIZA++ toolkit that was used as the word aligner, to create a word alignment by learning the probabilistic lexicon from the corpus using the IBM models (vd. Chapter 2).

I started the training by instantiating the process with the script `train-model.perl` that executes all of necessary steps. I will now describe these sequential steps that were undertaken to accomplish the training of the SMT component:

1. Data preparation: The parallel corpora was automatically converted into a format that is suitable for the GIZA++ toolkit. Vocabulary files were created and the parallel corpora converted into a numberized format using the `plain2snt.out` tool from the GIZA++ toolkit.

   For each language a vocabulary file was created containing a word identifier followed by the word itself and the word count, as for example:

   ```
   1    the  4800171
   2    .    4138634
   3    ,    3324992
   4    of   2757714
   5    to   2112111
   6    and  1859419
   7    in   1594733
   8    a    1021882
   9    is   938977
   10   that 857343
   ...
   7219 elbow 631
   ```

   After this, each original sentence was converted into three lines with a numberized format. The first line representing the sentence frequency (typically with the value 1, although it could be used to apply weights between sentences), the second and third line representing the words identifier for the source and target language. As for example when representing the phrase *the elbow*, with Portuguese as the source language and English as the target language one could find the following representation:

   ```
   1
   6 8921
   1 7219
   ```

   In the third line, the sequence $1\ 7219$ represents the identifier of each word of the phrase *the elbow*, where the word *the* is identified with the value 1 as seen in the vocabulary file above and the word *elbow* with the identifier 7219. The second line represents the identifiers for the Portuguese phrase.

2. Word-Alignment with Giza++: Each of the target language words was automatically aligned to a source language word and vice versa by Giza++. From this step,

two alignment files result which are composed by three lines for each sentence pair: the first line stating for each sentence the source and target word length along with the alignment score, the second line containing the source sentence in its surface form and the third line containing the target sentence with the alignment reference to the source word. As for example, the English source sentence *My arm hurts* which has in the parallel corpora the corresponding Portuguese sentence *Dói-me o braço* could result in the following representation:

```
#Sentence pair (93326) source length 3 target length 4
                          alignment score : 0.000516087
My arm hurts
NULL({}) Dói ({3}) me ({1})  o ({}) braço ({2})
```

The special token NULL is used in each target sentence to store the source words without alignment.

All the word alignments were learned through Giza++ by applying the IBM model, which in turn uses the Expectation Maximization (EM) algorithm to find the most probable word alignments. The EM algorithm begins by initializing for each word a set of word alignments with all possible words, equally likely, that is, for each source word, a probability table is collected, containing the alignment probability to every possible target word. Then by iteration and taking into account the frequency of each source and target word, the probabilities mass is adjusted, converging to the most probable alignment.

Some problems occurred in this step, mainly with the memory requirements. To cope with this problem the `snt2cooc` program was used. The `snt2cooc` collects a list of words that may co-occur prior to the running of GIZA++. It was configured to run with a sort buffer size of 200M and compressing with gzip.

3. Alignment of words: The resulting Giza++ alignments were intersected using the default heuristic, the `grow-diag-final`, applied with the `giza2bal.pl` script. Although the previous step used a IBM model that can explicitly take one-to-many alignments into account, the translation probabilities $p(s|t)$ are not yet adapted to the phrase-based model.

   In this step a symmetrization of the IBM model alignments occurs, allowing to align multiple input words to an output word by intersecting the two starting alignments and by adding gradually the additional alignment points. By taking into account the target to source alignment and the source to target alignment, the `grow-diag-final` heuristic adds the alignment points that intersect and also the diagonal neighborhood words, the ones that not fully intersect but correspond to one-to-many alignments.

The generated file is a numberized corpus containing the alignment between the source and target language for each word. As for example, given the source language sentences,

```
My arm hurts
Help me please
```

and the target language sentences,

```
Dói -me o braço
Ajuda -me por favor
```

it could be generated the following alignment file line:

```
1-2 1-3 2-4 3-1
1-1 2-2 3-3 3-4
```

Where in this case the word *My* (1) was aligned with *-me* (2) and *o* (3), represented as *1-2 1-3*. Some bad alignments are bound to happen, like the *please* (3) with the *por* (3).

4. Creation of the lexical translation table: This table was automatically created by using a maximum likelihood estimation. It was done with the `train-model.perl` script. For both languages, a file is created containing the source word and target word followed by the word lexical probability, as for example the source word *injuries* could have the following lexical translation probabilities:

```
injuries    lesões          0.914
injuries    escoriações     0.004522
injuries    sofrem          0.000321
injuries    sinistralidade  0.0279
injuries    danos           0.0158
...
```

5. Extraction of phrases: All the phrases were gathered, using the default maximum phrase length of 7 words, and consolidated into a unique file. This step is performed through the `train-mode.perl`. An inverted alignment was also generated. Following the previous examples, one would find, for example, the following entry in the generated alignment file:

```
my arm hurts ||| dói -me o braço ||| 1-2 1-3 2-4 3-1
my arm |||  meu braço ||| 1-1 2-2
my ||| meu ||| 1-1
...
arm hurts ||| dor no braço ||| 1-2 1-3 2-1
```

6. Generation of phrase scores: A translation table was automatically created from the stored phrases translation pairs using the `train-model.perl` script. In this phase the inverse and direct translation probability are estimated along with two additional and optional scores, the direct lexical weighting and phrase penalty.

   The phrase translation probability $p(s|t)$ is estimated by sorting the extracted phrases and collecting counts and compute the probability for each target phrase, the same process is applied to estimate the inverse phrase translation probability $p(t|s)$.

   The lexical weighting is a smoothing method that takes into account the relative frequency of same word alignments; the phrase penalty is set to the constant value of $exp(1) = 2.718$.

The resulting phrase tables were too large to fit into memory so binary phrase tables were created. These allowed to load into memory only the segments of the phrase table necessary to translate a sentence. The `processPhraseTable` tool from Moses was used for this purpose.

Tuning

The tuning consists in optimizing the decoder search parameters so that a better translation can be obtained in future translation searches.

The Moses decoding of the scores for translation hypotheses was done using the available linear model. The linear model included features such as the language models, phrase tables, phrase and word counts. For tuning the decoders, the optimal weights of these features were searched for the linear model, being the optimal weights those that maximize the translation performance using the BLEU metric for evaluation.

I set up the tuning process using `MERT`, a Batch algorithm implementation. The process consisted in the automatic decoding of all sentences in the tuning set, using the ACL corpora (vd. Section 4.3, "Preparation of corpora") and the trained model described above. While decoding, it generated an n-best list of scores, comparing with the reference translation and using the BLEU metric. With the resulting BLEU, score the model weights are updated. The process is repeated with the new weights until the convergence criterion is satisfied, this is, no gain is obtained for any parameter weight update.

4.4 Text-to-Speech Synthesis

The MSP TTS uses state of the art HMM-based speech synthesis. It applies a set of front-end modules such as Sentence Separator, Word Splitter, Text Normalization and Homograph Disambiguation modules.

The MSP TTS system is a ready to use TTS where the main components cannot be accessed. It works by taking a text string as input and producing an audio stream as output.

The implementation was done by programming, in C#, an interface to the TTS system. The main configuration was done in the dynamic choice of the language of the input string.

The language pack used for American English was the MSSpeech SR en-US TELE, and the MSSpeech SR pt-PT TELE for European Portuguese. These language packs are made available by MS and they are closed to configuration.

After initializing a `SpeechSynthesizer` instance from the MSP TTS, the language pack is chosen and some minor parameterization is done. The parameterization includes:

- Speech rate, with a range of values from -10 to 10, from a slower to higher speech rate: It was set to 1.
- Volume, with a range of values from 0 to 100, from silence to loudest: It was set to 100.

All the speech output is done to a WAV stream for its transmission to be possible through web services.

4.5 Web Services

The implementation of the web services for the ASR and TTS components was undertaken in a Windows environment, by using the ASP.net framework with C# as source language and running it in the Internet Information Services (IIS).

The ASR web service has the following arguments:

- Audio: An audio file in byte format, containing the speech to be recognized.
- Language code: a string that references the speech language. Depending on the language, a different language model and speech engine are used.

The web service returns a string containing the result of speech recognition, which is the phrase with the higher confidence score.

The TTS web service, in turn, has the following arguments:

- Sentence: The text to be synthesized to speech.
- Language code: A string that references the language in which the speech is to be synthesized.

Upon receiving a request, the web service permits the interfaces with the TTS Engine, which returns the resulting synthesized audio. The audio is streamed to memory and returned by the web service as an audio file in a byte format.

The implementation of the web services for the SMT system was undertaken in a UNIX environment, using the NuSOAP Toolkit for PHP and running in an Apache Server.

The SMT web service has the following arguments:

- Sentence: The source text to be translated.
- Source language: A string representing the code that allows to identify the language of the source phrase.
- Target language: A string representing the code that allows to identify the target language to which the source phrase should be translated.

The web service upon receiving a request, communicates to a server daemon to obtain the translation. The daemon contains an instance of a Moses running decoder. Two daemons and two decoders are used, one for each translation direction.

After the response from the decoder, the web service returns a string containing the translation.

4.6 Web Application (LX-Translator)

The LX-Translator was developed in HTML, Javascript and PHP. It allows the use of S2ST from English to Portuguese and Portuguese to English in the Google Chrome web browser. The SMT and TTS functionalities make use of the implemented web services. The ASR uses the Google Chrome implementation of the W3C Web Speech API Specification.

The speech recognition language is automatically adjusted for the selected source language.

A snapshot of the user interface can be seen in Figure 4.6. The demo of this application is available at: `http://lxtranslator.di.fc.ul.pt`

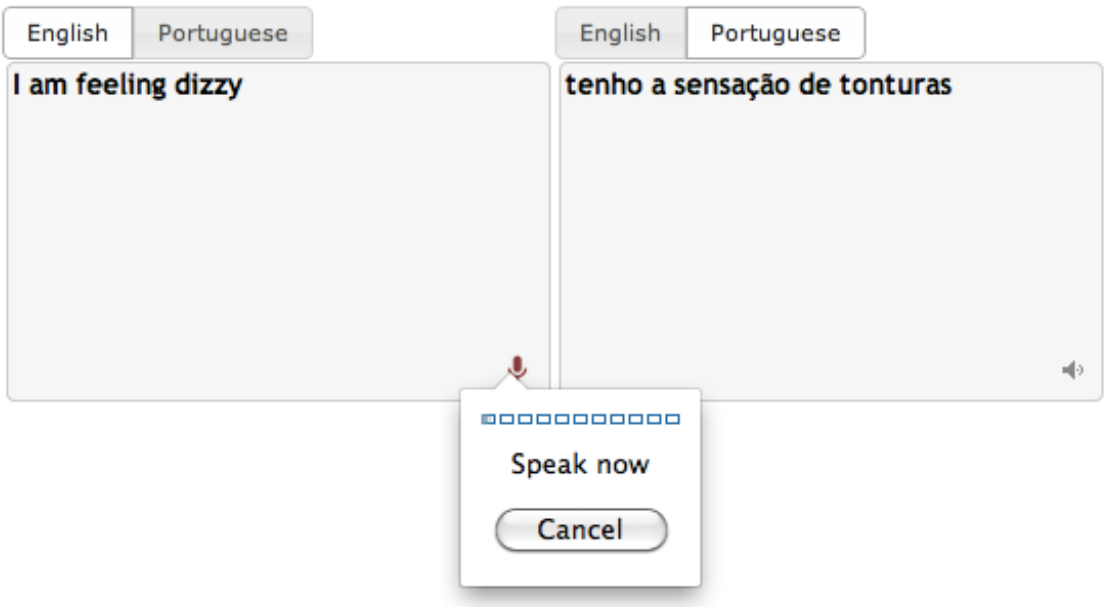

Figure 4.6: LX-Translator interface

4.7 Mobile Application (LX-iTranslator)

The LX-iTranslator was developed using objective-C over the iOS framework, running on IPhone smartphones from Apple.

The LX-iTranslator supports the speech-to-speech dialogue between two persons using the English and Portuguese languages.

The iPhone device is expected to stand in between two persons so that each of them can use his half of the screen.

Each person access a button that, when pressed, waits for the user to speak, recording his speech. The moment the button is no longer pressed the recorded audio is used as input for the S2ST system. Each person also has access to a text box where the speech recognition and the translations results are presented. The synthesized speech is heard upon it is received in the device, being possible to turn it off in the application.

The LX-iTranslator works along the following steps:

- After the audio is recorded, it is converted to a byte encoded format. A request is then made to the ASR web service, where the byte encoded object and its source language code identification are passed as arguments. The application then waits for the ASR web service response.

- On the arrival of the ASR web service response, the best phrase is presented in the sender text box while it is sent to the SMT web service along with its source and target language code identification. The application then waits for the SMT web service response.

- On the arrival of the SMT web service response, the resulting translation string is shown in the receiver text box. With this string, a request is made to the TTS web service, using the target language code as an argument.

- On the arrival of the TTS web service response, the resulting synthesization is converted from byte encoding to wave form and played as audio in the speakers of the device.

A snapshot of the LX-iTranslator interface can be seen in Figure 4.7.

Due to the necessary license, fees and approval from Apple to publish a mobile application and the fact that the application is just a prototype not yet ready for the public domain, the mobile application was not publish through the Apple Application Store (digital distribution platform for the iOS mobile applications). An iPhone with the LX-iTranslator installed prototype can be requested at the NLX-Natural Language and Speech Group at the University of Lisbon, Department of Informatics.

Figure 4.7: LX-iTranslator mobile interface

Chapter 5

Evaluation

The evaluation of the S2ST system was done by evaluating the ASR component and the SMT component.

The evaluation of the TTS was not performed on the one hand, due to the lack of resources to gather a group of human testers and organize a rigorous evaluation exercise, on the other hand, and above all, because I would not be evaluating, in this respect, a substantial result of my work since the TTS component used is closed to preparation or parameterization.

5.1 Automatic Speech Recognition

The evaluation of the ASR system was done using the WER metric.

From the Europarl corpus I extracted and recorded a speech corpus, the Med-Speech Corpus, in order to be used as a test set for the ASR system. The remaining corpus was used for the creation of the ASR component language model. The Med-Speech Corpus transcription consists of 200 spoken sentences, with a total of 3911 words, with approximately 20 words per sentence, and it can be seen in Appendix D. Due to the lack of a native English person to record the spoken sentences, the Med-Speech Corpus was only available for the Portuguese language.

An off-the-shelf headset microphone was used for the recording.

The resulting speech recognition was normalized and then evaluated with the WER metric using the Apertium toolbox that provides a WER evaluator.

Using the remaining EUMed Corpora (vd. section 4.1), a set of different size language models were evaluated, all of them using the Portuguese language. Each language model was created from the previous language model corpus by extending it with extra 500,000 sentences.

Upon the ASR system recognition of a spoken sentence, if more than one hypothesis occurred, only the one with the highest confident score was selected for evaluation. The resulting evaluation can be seen in Figure 5.1. The greater the size of the corpus used in

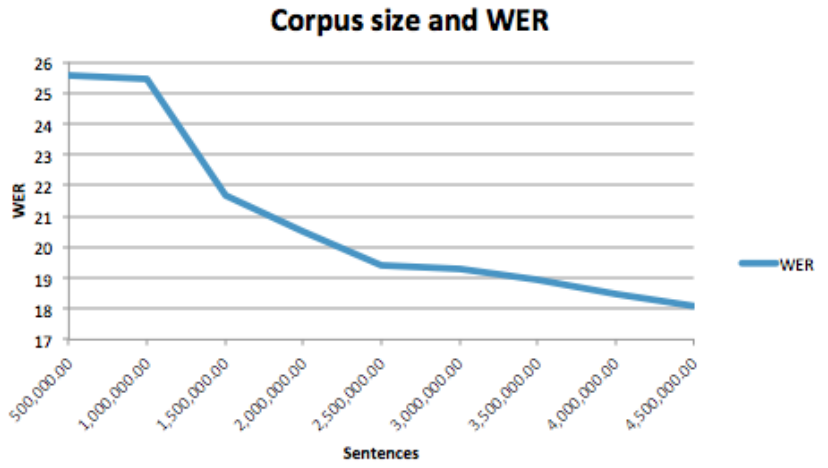

Figure 5.1: Word Error Rate by Corpus size

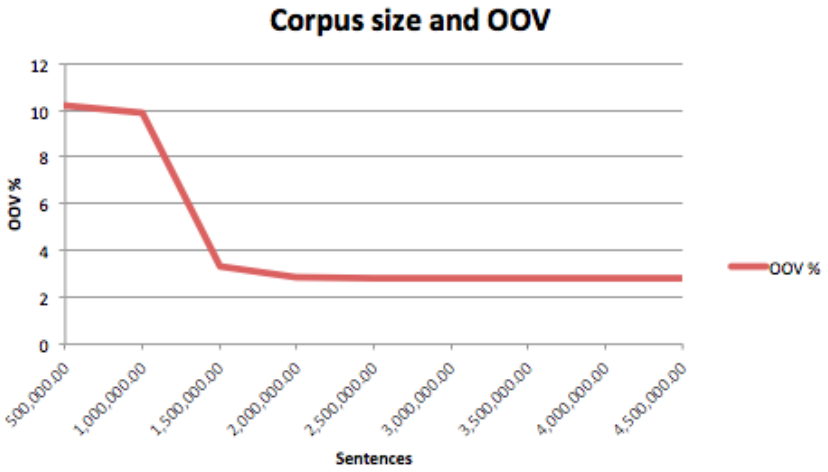

Figure 5.2: Out-of-vocabulary Words by Corpus size

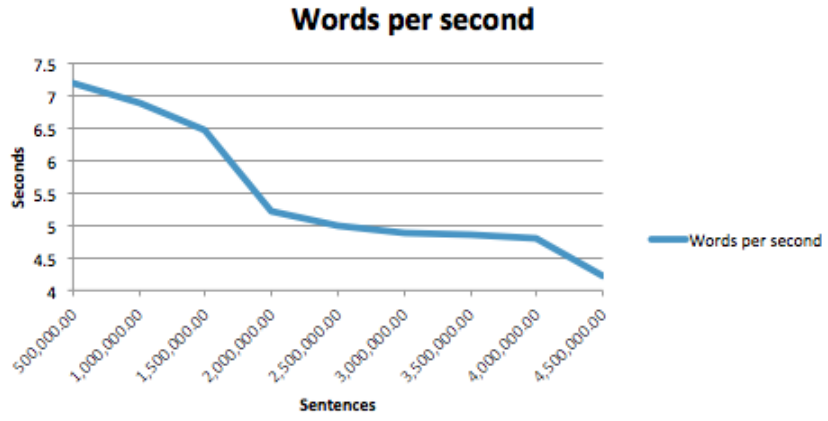

Figure 5.3: Words recognized per second

the language model, the greater is the speech recognition success. This can be noted by the reduction of the WER.

In order to try to explain the most accentuated reduction (3.76) in the WER evaluation from the 1,000,000 to the 1,500,000 sentences language model, the percentage of the out-of-vocabulary words was calculated, since the out-of-vocabulary words have a great influence in the speech recognition results. As it can be seen in Figure 5.2, the improvement of the recognition is related to the reduction of the out-of-vocabulary words.

The sudden drop in the out-of-vocabulary words is due to the topic covered by the train set and the test set. The domain of the training corpus is varied, however the main topic of the test set, environment problems, occurs in the train set in the portion added when the training set is incremented from 1.000.000 and 1.500.000 sentences.

A completely fair comparison with the state of the art is not viable due to the impossibility of training the acoustic model. Nevertheless, my results can still be put into perspective taking into account the results of the International Workshop on Spoken Language Translation (IWSLT) from 2012.

The goal of the IWSLT 2012 ASR campaign, for the speech recognition evaluation, was to transcribe English recordings from TED conferences. The data set used had 1,124 sentences with a 2,800 word vocabulary, a higher number of sentences and word vocabulary in comparison with this work (200 sentences, 1385 word vocabulary). The best system in the campaign achieved a WER score of 12% [59] [60], to which the WER score of 18% compares pretty competitively, given the circunstances.

Since the increase in the size the language model induces a longer recognition time, an experiment speech recognition time was conducted. The results can be seen in Figure 5.3. As expected, the number of words recognized per second decreased for larger language models.

The language model generated with 4,500,000 sentences was the chosen language model, obtaining the best evaluation and with a still acceptable average of 4 words per second for practical usage.

5.2 Statistical Machine Translation

For the SMT evaluation, the ACL Third Workshop on Statistical Machine Translation (2008) [58] test set was used.

Six versions of the SMT component were created and compared, where three main approaches were used for the Portuguese-English pairs. The first version was obtained by training a phrase-based model using only the Europarl corpus. The second version was done by also using a phrase-based model with the EUMed Corpora for its training

(without the Part-of-speech and Lemma features). A third version used a factored model with the EUMed Corpora annotation for its training.

Before choosing a factored model, a set of factored models was experimented with. This is due to the possible parameterization of the order and factors to be used for the translation and generation. For example, one can create a model in which the source language Lemma factor is translated to the target language and from that translation it could be generated the target language surface words. With a large corpora, as the one that was used, some of the models were not feasible for its practical use in a S2ST component. Some of them would take more than half an hour to produce a translation and others would have a large memory requirement that was not available. These situations occurred in models that use a more rich factor parameterization for the translation and generation.

The factored model that appeared to be more practically viable is the one that I used for the comparison of the systems. For the language model, it uses the surface form along with a part-of-speech language model. For the translation model, the translation was done with the following sequence of mapping steps:

- 1) Translation of input surface form to output surface form.
- 2) Translation of input part-of-speech form to output part-of-speech form.
- 3) Translation of surface forms given the output surface form and part-of-speech factors.

The comparison of the systems in terms of BLEU scoring can be seen in Figure 5.4.

In all systems, the translation from Portuguese to English obtained better results as expected, since the translation occurs from a richer morphological language to a poorer one.

The scores are within an acceptable range when comparing with the 2008 ACL Workshop on Parallel Texts [61], where using the same test set, it was obtained a 0.322 and 0.294 BLEU score, for the Portuguese to English and English to Portuguese translation respectively, versus the 0.2867 and 0.2702 BLEU score obtained in the phrase-based model I implemented.

The comparison was done with the Official Europarl test set from WMT 2008 ACL Workshop due to the available results from the EuroMatrix Evaluation (`http://matrix.statmt.org/matrix`) for the English to Portuguese and vice-versa results.

As expected, using a larger corpora yielded better results. To establish if the use of a larger training corpus would also bring better results when applied to different test corpora, from different domains, like the medical-patient dialogue domain, another evaluation exercise was performed. This evaluation used sets of 200 sentences from each of these four different corpus:

- EU BookShop (EUB), a parallel corpus from book brochures.

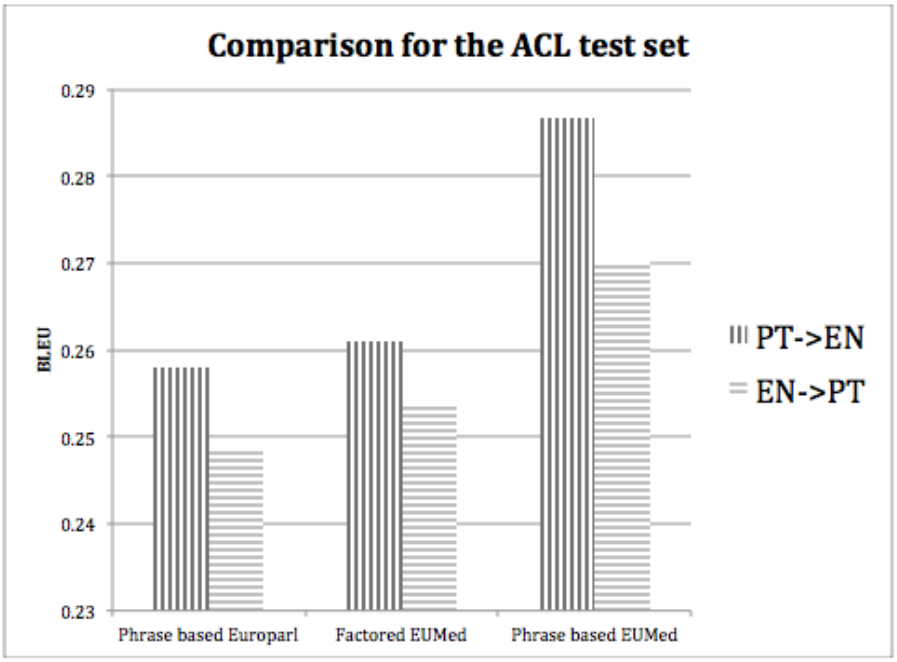

Figure 5.4: Comparison of three SMT systems created in terms of BLEU Scoring

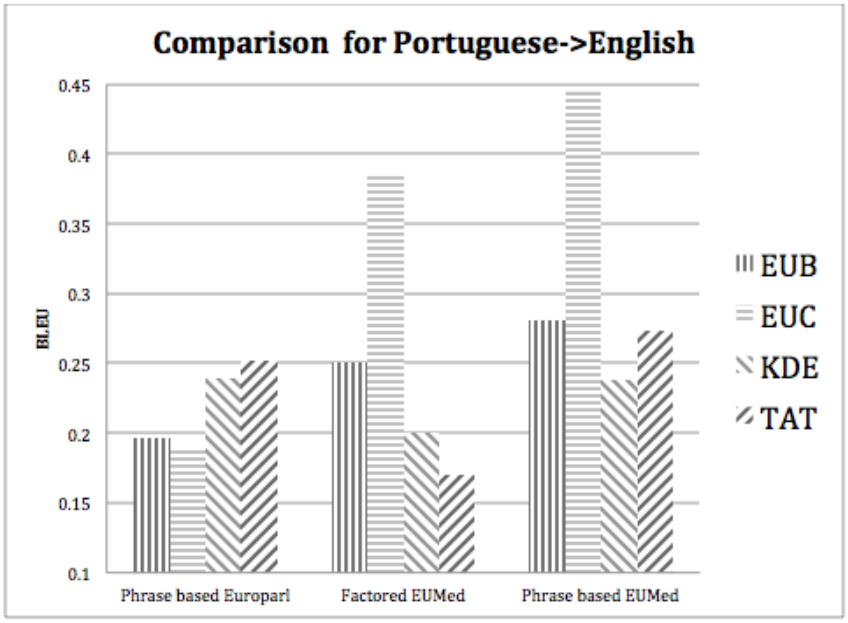

Figure 5.5: Comparison of the BLEU scores for the versions with the Portuguese to English direction applied to the EU BookShop, EUconst, KDE and Tatoeba test corpora

- EUconst (EUC), a parallel corpus collected from the European Constitution.
- KDE, a parallel corpus of KDE4 localization files.
- Tatoeba (TAT), a collection of translated sentences form Tatoeba, a large database of example sentences translated.

The main reason for the choice of these four different corpus was the difference in the domain and their availability [62]. These corpora are described in more detail in Appendix F.

In Figures 5.5 and 5.6 it can be seen the Portuguese to English and the English to Portuguese translation evaluation respectively.

In general, both the Factored model as well as the Phrase based model using the EUMed Corpora surpassed the Phrase based model using only the Europarl corpus. This seems to indicate a better capacity of generalization, that offer prospects of better performance in domains that are different from the training domain, such as, for instance, the doctor-patient dialogues, that are of interest here.

Given these results, for these essays, the SMT component was eventually implemented using the phrase-based model created with the EUMed Corpora as the SMT system.

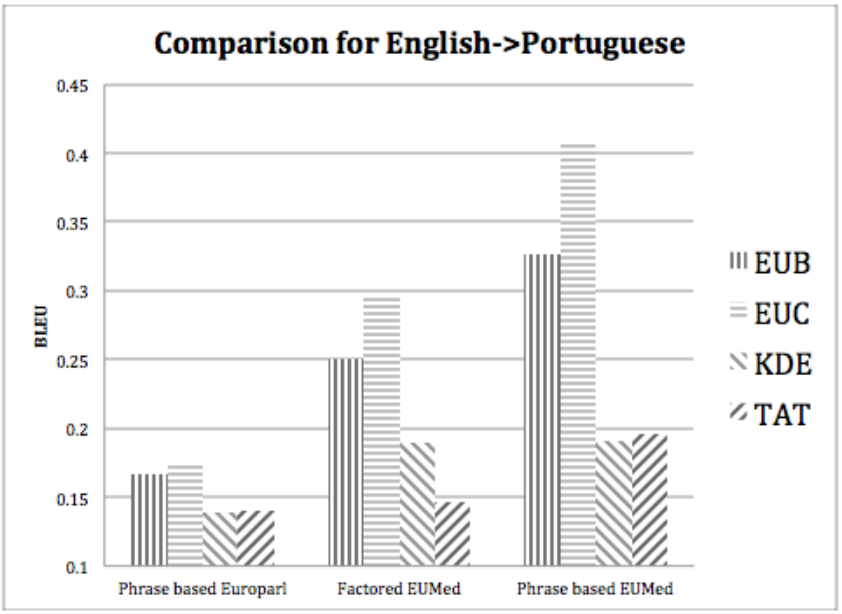

Figure 5.6: Comparison of the BLEU scores for the versions with the English to Portuguese direction applied to the EU BookShop, EUconst, KDE and Tatoeba test corpora

Chapter 6

Conclusions and Future work

6.1 Introduction

In this work, a speech-to-speech translation prototype system was built aiming for a doctor-patient dialogue domain and supporting the Portuguese and English languages. The system was made available in two forms of application, web and mobile.

The system consisted of three components: automatic speech recognition, statistical machine translation and text-to-speech synthesis.

The automatic speech recognition system allows the recognition of speech from both languages. This was done by implementing a speech engine trained with language models created specifically for the task. The creation of these language models was done by experimenting with context-free-grammars and a statistical language model, where the latter was chosen.

The statistical machine translation component allows the translation of the output from the automatic speech recognition system. The phrase-based and factored models were used for the creation of versions of the machine translation component. Different corpora were used for the training and evaluation of the systems, some of which needed a more detailed preparation, as those annotated with lemma and part-of-speech for the factored models. The final choice fell over a phrase-based model with the larger corpora gathered.

Since the available technology for the text-to-speech component allowed it to be deployed without much implementation work, the focus of this work went to the automatic speech recognition and, above all, statistical machine translation components.

The pipelining of the components was made through the use of web-services.

Since the components used statistical approaches which require large corpora, this work also describes the gathering and preparation of corpora necessary for the training of the automatic speech recognition and statistical machine translation components.

The evaluation of the automatic speech recognition and statistical machine translation showed that the performance of the system developed compares very competitively with the state of the art.

With this work, it was demonstrated that the prototyping of a speech-to-speech translation system for translation pairs with Portuguese and English from available practical technologies solutions is a practically viable task nowadays and definitely a future available tool world-wide.

6.2 Conclusions

One of the most important endeavors for the viability of this project was the gathering of parallel corpora. Since all the components methods used in this project were statistical, the necessity of data was crucial and as expected the more the merrier.

The gathering of parallel corpora proved to be more demanding than I expected, clearly indicating the need of a special attention on the necessary data for the training of the three components. The parallel corpora need to be aplenty, with the least possible noise and ideally in the domain which the translation system is targeting. Something that by the time of this project did not exist, taking into account that the domain of the project is the medical interview. The lack of parallel corpora in the ideal domain led to search and gather corpora that were acceptable but restrain the success of the S2ST system.

I think it is also important to point out that there are limitations on the availability of corpus for other languages, different from Portuguese or English, which could be of a greater lost since the linguistic barrier with least accessible languages would have a greater necessity of an S2ST system.

In the ASR and TTS components, I noticed that there is still a small step to be made for the existence of an open source solution that enable the use of Portuguese, in particular at the level of acoustic data necessary for the training of the components.

In the SMT component, Moses proved to be an excellent open source solution, allowing for a dynamic and detailed implementation of a SMT system.

Regarding the S2ST system as a whole, there are no all in one solutions at the moment. These separation has an effect on the evaluation task. Since the ASR, SMT and TTS research is done separately, there were no common baseline and evaluation corpus that could be used for the training and evaluation of all the system within the same domain.

Although an all in one S2ST would be ideal, the integration through web services remains important, since nowadays the most viable way of using a S2ST is with a mobile phone, given their portability but, their computational power is still not yet ready to integrate a full S2ST system.

Finally, acknowledging the task of speech recognition, translation and synthesization as a cognitive and knowledgeable demanding task for a human, one may not forget how fascinating it is that nowadays one can perform such task with the help of computer science and especially natural language processing.

6.3 Future work

The future work would focus on the creation of ASR and TTS components from the ground up or with an open-source solution, allowing a wider customization, fine-tuning and optimization of the system, including the creation of acoustic models, and the possible comparison with state of the art systems.

The development of complex factored models in the SMT component may yield better results for the translation, along with the use of hybrid machine translation systems.

The gathering of large volumes of medic-patient dialogues, which includes human transcription and translation would have the most influential impact in an S2ST system for this domain, which I presume would bring a much more precise speech recognition and translation.

Appendix A

Anamnesis Corpus

Olá
Bom dia
Boa tarde
Boa noite
Tchau
Como vai
Como vão as coisas
Quais são as novidades
Como tem passado
Adeus
As melhoras
Prazer em conhecê-lo
Eu confio que tudo está bem
Bons olhos te vejam
Como estão as coisas
Como é que vai
Como é que a vida o tem tratado
Prazer em conhecê-lo
O meu nome é
Eu gostaria de me apresentar
Eu não acho que nós já nos conhecemos
Olá você deve ser a senhora
Olá você deve ser o senhor
Olá deve ser o senhor
Eu sou o
Já nos conhecemos
Eu sou a
O que o traz cá

O que é que se passa
Porque veio cá hoje
Qual é o seu problema
O que lhe aconteceu
O que se passa consigo
O que tem de mal
Em que posso ajudar
Como o posso ajudar
Dói-me o braço
Tenho uma dor no braço
Magoei-me no braço
Não aguento com dores no braço
O meu braço dói-me
Uma dor no braço
Acidente de carro
Aleijei-me no braço
Tive um acidente de carro
Dores no braço
Vim por causa de dores
Não consigo mexer o braço
Tenho o braço dolorido
Tenho o braço dorido
Em que braço
De que lado do braço
Em que parte do braço
Mostre-me onde se aleijou
Cortou-se
Como foi o acidente
Como ocorreu o acidente
Magoou-se em mais algum sitio
A dor irradia para algum lado
A dor irradia para algum sitio
Ficou inchado
Esquerdo ou direito
Direito ou esquerdo
Qual deles
Aonde
No cotovelo
No antebraço

Só no braço
Consegue mover o braço
Esquerdo
Parte esquerda
Não me cortei
Só no braço
Não irradia
Ficou inchado
No cotovelo
Mal
Pouco
Dói ao mover
Dói quando mexo
Um pouco
Isto dói
Dói assim
Dobre o cotovelo
Tente dobrar
Pode dobrar o cotovelo
Tente dobrar o cotovelo
Relaxe
Deixe-me ser eu a dobrar
Deixe-me fazer
Deixe-me dobrar
Não faça força
Mexa os dedos
Aperte os meus dedos
Sente igual nas duas mãos
Sensibilidade idêntica nas duas mãos
Pode ter
Pode ser que tenha
Talvez tenha
Cotovelo partido
Precisa de fazer uma radiografia
Vai ser necessária uma radiografia
Não tem nada partido
Tem o braço partido Poderá ter o braço partido Parece-me partido Tem apenas o cotovelo magoado
Deve manter o braço imobilizado

Pode colocar gelo no cotovelo
Vou-lhe prescrever um analgésico
Sim
Não
Talvez
Por favor
Se faz favor
Faça favor
Obrigado
Não percebi
Repita
Que
De
Um
Para
Se
Me
Uma
Está
Com
Do
Por
Te
Os
Bem
Em
Ele
Isso
Mas
Da
Como
No
As
Mais
Na
Meu
Aqui
Muito
Foi

Estou
Vamos
Ela
Fazer
Vai
Isto
Já
Tem
Só
Minha
Nos
Tudo
Ao

Appendix B

Med-CFG

```
frase --> sn,sv.
frase --> sv.
frase --> sn.
frase --> sa.
frase --> s_int,frase.
fbarra --> comp,frase.

sn --> sn,sn.
sn --> det,nbarra.
sn --> nbarra.

nbarra --> nome.
nbarra --> nbarra,adjectivo.
nbarra --> adjectivo,nbarra.
nbarra --> nbarra,sp.
nbarra --> poss,nbarra.

sp-->prep,sn.
sp-->prep,sv.
sp-->contr,nbarra.

sv --> verbo.
sv --> verbo,sn.
sv --> verbo,adjectivo.
sv --> sv,adverbio.
sv --> adverbio,sv.
sv --> verbo, fbarra.
sv --> sv,sp.
sv --> sp,sv.

det --> [a].
det --> [o].
det --> [as].
det --> [os].
det --> [algum].

s_int --> sp_int.
s_int --> adv_int.
s_int --> sn_int.
```

```
sp_int -> prep,sn_int.
sn_int --> det_int, nbarra.
sn_int --> pron_int.

poss --> [meu].
poss --> [seu].

contr--> [no].
comp --> [que].
det_int --> [que].
adv_int --> [como].

pron_int --> [quais].
pron_int --> [qual].
pron_int --> [o,que].
pron_int --> [que].

verbo --> [acho].
verbo --> [aconteceu].
verbo --> [ajudar].
verbo --> [aleijei].
verbo --> [aperte].
verbo --> [aleijou].
verbo --> [apresentar].
verbo --> [colocar].
verbo --> [conhecemos].
verbo --> [consegue].
verbo --> [consigo].
verbo --> [cortei].
verbo --> [deixe].
verbo --> [deve].
verbo --> [dobrar].
verbo --> [dobre].
verbo --> [doi].
verbo --> [doi-me].
verbo --> [estou].
verbo --> [esta].
verbo --> [estao].
verbo --> [faz].
verbo --> [fazer].
verbo --> [faca].
verbo --> [ficou].
verbo --> [foi].
verbo --> [gostaria].
verbo --> [magoei].
verbo --> [magoou].
verbo --> [manter].
verbo --> [mexa].
verbo --> [mexer].
verbo --> [mostre].
verbo --> [mover].
verbo --> [ocorreu].
verbo --> [passa].
verbo --> [percebi].
verbo --> [pode].
verbo --> [posso].
```

```
verbo --> [precisa].
verbo --> [prescrever].
verbo --> [ser].
verbo --> [sou].
verbo --> [sao].
verbo --> [tem].
verbo --> [tenha].
verbo --> [tenho].
verbo --> [tente].
verbo --> [ter].
verbo --> [tive].
verbo --> [traz].
verbo --> [vai].
verbo --> [vamos].
verbo --> [veio].
verbo --> [vejam].
verbo --> [vim].
verbo --> [vou].
verbo --> [vao].
verbo --> [confio].
verbo --> [irradia].

verbo --> [mexo].
verbo --> [relaxe].
verbo --> [repita].

nome --> [acidente].
nome --> [melhoras].
nome --> [aguento].
nome --> [antebraco].
nome --> [braco].
nome --> [carro].
nome --> [causa].
nome --> [coisas].
nome --> [cotovelo].
nome --> [dia].
nome --> [dedos].
nome --> [dor].
nome --> [dores].
nome --> [favor].
nome --> [forca].
nome --> [gelo].
nome --> [lado].
nome --> [maos].
nome --> [noite].
nome --> [nome].
nome --> [novidade].
nome --> [olhos].
nome --> [parte].
nome --> [partido].
nome --> [radiografia].
nome --> [problema].
nome --> [prazer].
nome --> [senhor].
nome --> [senhora].
nome --> [sensibilidade].
```

```
nome --> [sitio].
nome --> [tarde].
nome --> [vida].
nome --> [analgesico].

prep --> [em].
prep --> [de].
prep --> [por,causa,de].

conj --> [como].
conj --> [mas].
conj --> [ou].
conj --> [porque].
conj --> [quando].
conj --> [que].

adverbio --> [hoje].
adverbio --> [tarde].
adverbio --> [nao].

adjectivo --> [direito].
adjectivo --> [esquerdo].
adjectivo --> [esquerda].
adjectivo --> [belo].
adjectivo --> [boa].
adjectivo --> [bom].
adjectivo --> [bons].
adjectivo --> [identica].
adjectivo --> [igual].
adjectivo --> [necessario].
adjectivo --> [dolorido].
adjectivo --> [dorido].
adjectivo --> [imobilizado].
adjectivo --> [inchado].
adjectivo --> [magoado].
adjectivo --> [passado].
adjectivo --> [tratado].

sa --> [adeus].
sa --> [obrigado].
sa --> [tchau].
sa --> [ola].
```

Appendix C

Med-SRGS

```
<?xml version="1.0"encoding="UTF-8" ?>
<grammar version="1.0" xml:lang="pt-PT" mode="voice" root="grammar"
xmlns="http://www.w3.org/2001/06/grammar" tag-format="semantics/1.0">

<rule id="grammar" scope="public">
  <item><ruleref uri="#Frase"/></item>
</rule>

<rule id="Frase">
  <one-of>
    <itemrepeat="0-1"><ruleref uri="#Frase1"/></item>
    <itemrepeat="0-1"><ruleref uri="#Frase2"/></item>
    <itemrepeat="0-1"><ruleref uri="#Frase3"/></item>
    <itemrepeat="0-1"><ruleref uri="#Frase4"/></item>
  </one-of>
</rule>

<!-- fbarra -> comp,frase. -->
<rule id="Fbarra">
  <item>que</item>
  <item><ruleref uri="#Frase"/></item>
</rule>

<!-- frase -> sn,sv -->
<rule id="Frase1">
  <itemrepeat="0-1"><ruleref uri="#SN"/></item>
  <itemrepeat="0-1"><ruleref uri="#SV"/></item>
</rule>

<!-- frase -> sn -->
<rule id="Frase2">
  <item><ruleref uri="#SN"/></item>
</rule>

<!-- frase -> sa -->
<rule id="Frase3">
  <item><ruleref uri="#sa"/></item>
</rule>

<rule id="sa">
```

```
  <one-of>
    <item>adeus</item>
    <item>obrigado</item>
    <item>tchau</item>
    <item>ola</item>
  </one-of>
</rule>

<!-- frase -> s_int,frase -->
<rule id="Frase4">
  <item><ruleref uri="#s_int"/></item>
  <item><ruleref uri="#Frase"/></item>
</rule>

<rule id="s_int">
  <one-of>
    <item><ruleref uri="#s_int1"/></item>
    <item><ruleref uri="#s_int2"/></item>
    <item><ruleref uri="#s_int3"/></item>
  </one-of>
  <item><ruleref uri="#Frase"/></item>
</rule>

<!-- sp_int -> prep,sn_int -->
<rule id="s_int1">
  <item><ruleref uri="#prep"/></item>
  <item><ruleref uri="#sn_int"/></item>
</rule>

<rule id="prep">
  <one-of>
    <item>em</item>
    <item>de</item>
    <item>por causa de</item>
  </one-of>
</rule>

<rule id="sn_int">
  <one-of>
    <item><ruleref uri="#sn_int1"/></item>
    <item><ruleref uri="#sn_int2"/></item>
  </one-of>
</rule>

<!-- sn_int -> det_int, nbarra -->
<rule id="sn_int1">
  <item><ruleref uri="#det_int"/></item>
  <item><ruleref uri="#NBARRA"/></item>
</rule>

<rule id="det_int">
<item>que</item>
</rule>

<!-- sn_int -> pron_int -->
<rule id="sn_int2">
```

```
  <item><ruleref uri="#pron_int"/></item>
</rule>

<rule id="pron_int">
  <one-of>
    <item>quais</item>
    <item>qual</item>
    <item>o que</item>
    <item>que</item>
  </one-of>
</rule>

<!-- s_int -> adv_int -->
<rule id="s_int2">
  <item>como</item>
</rule>

<!-- s_int -> sn_int -->
<rule id="s_int3">
  <item><ruleref uri="#sn_int"/></item>
</rule>

<rule id="SN">
  <item><ruleref special="GARBAGE"/></item>
  <one-of>
    <item><ruleref uri="#SN1"/></item>
    <item><ruleref uri="#SN2"/></item>
    <item><ruleref uri="#SN3"/></item>
  </one-of>
</rule>

<!-- sn -> sn,sn -->
<rule id="SN1">
  <one-of>
    <itemrepeat="0-2"><ruleref uri="#SN2"/></item>
    <itemrepeat="0-2"><ruleref uri="#SN3"/></item>
  </one-of>
</rule>

<!-- sn -> det,nbarra -->
<rule id="SN2">
  <item><ruleref uri="#det"/></item>
  <item><ruleref uri="#NBARRA"/></item>
</rule>

<!-- sn -> nbarra -->
<rule id="SN3">
  <item><ruleref special="GARBAGE"/></item>
  <item><ruleref uri="#NBARRA"/></item>
</rule>

<rule id="NBARRA">
  <item><ruleref special="GARBAGE"/></item>
  <one-of>
    <item><ruleref uri="#NBARRA1"/></item>
    <item><ruleref uri="#NBARRA2"/></item>
```

```
    <item><ruleref uri="#NBARRA3"/></item>
    <item><ruleref uri="#NBARRA4"/></item>
    <item><ruleref uri="#NBARRA5"/></item>
  </one-of>
</rule>

<!-- nbarra -> nome -->
<rule id="NBARRA1">
  <item><ruleref uri="#nome"/></item>
</rule>

<!-- nbarra -> nbarra,adjectivo -->
<rule id="NBARRA2">
  <item><ruleref special="GARBAGE"/></item>
  <!--<item><ruleref uri="#NBARRA"/></item>-->
  <item><ruleref uri="#adjectivo"/></item>
</rule>

<!-- nbarra -> adjectivo,nbarra -->
<rule id="NBARRA3">
  <item><ruleref uri="#adjectivo"/></item>
  <item><ruleref uri="#NBARRA"/></item>
</rule>

<!-- nbarra -> nbarra,sp -->
<rule id="NBARRA4">
  <item><ruleref special="GARBAGE"/></item>
  <!--<item><ruleref uri="#NBARRA"/></item>-->
  <item><ruleref uri="#SP"/></item>
</rule>

<rule id="SP">
  <one-of>
    <item><ruleref uri="#SP1"/></item>
    <item><ruleref uri="#SP2"/></item>
    <item><ruleref uri="#SP3"/></item>
  </one-of>
</rule>

<!-- sp -> prep,sn -->
<rule id="SP1">
  <item><ruleref uri="#prep"/></item>
  <item><ruleref uri="#SN"/></item>
</rule>

<!-- sp -> prep,sv -->
<rule id="SP2">
  <item><ruleref uri="#prep"/></item>
  <item><ruleref uri="#SN"/></item>
</rule>

<!-- sp -> contr,nbarra -->
<rule id="SP3">
  <item>no</item>
  <item><ruleref uri="#NBARRA"/></item>
</rule>
```

```
<!-- nbarra -> poss,nbarra -->
<rule id="NBARRA5">
  <one-of>
    <item>meu</item>
    <item>seu</item>
    <item>teu</item>
    <item>nosso</item>
  </one-of>
  <item><ruleref uri="#NBARRA"/></item>
</rule>

<rule id="SV">
  <item><ruleref special="GARBAGE"/></item>
  <one-of>
    <item><ruleref uri="#verbo"/></item>
    <item><ruleref uri="#SV2"/></item>
    <item><ruleref uri="#SV3"/></item>
    <item><ruleref uri="#SV4"/></item>
    <item><ruleref uri="#SV5"/></item>
    <item><ruleref uri="#SV6"/></item>
    <item><ruleref uri="#SV7"/></item>
    <item><ruleref uri="#SV8"/></item>
  </one-of>
</rule>

<!-- sv -> verbo,sn -->
<rule id="SV2">
  <item><ruleref uri="#verbo"/></item>
  <item><ruleref uri="#SN"/></item>
</rule>

<!-- sv -> verbo,adjectivo -->
<rule id="SV3">
  <item><ruleref uri="#verbo"/></item>
  <item><ruleref uri="#adjectivo"/></item>
</rule>

<!-- sv -> sv,adverbio -->
<rule id="SV4">
  <item><ruleref special="GARBAGE"/></item>
  <!--<item><ruleref uri="#SV"/></item>-->
  <item><ruleref uri="#adverbio"/></item>
</rule>

<rule id="adverbio">
  <one-of>
    <item>hoje</item>
    <item>tarde</item>
    <item>nao</item>
  </one-of>
</rule>

<!-- sv -> adverbio,sv -->
<rule id="SV5">
  <item><ruleref uri="#adverbio"/></item>
```

```
  <item><ruleref uri="#SV"/></item>
</rule>

<!-- sv -> verbo,fbarra -->
<rule id="SV6">
  <item><ruleref uri="#verbo"/></item>
  <item><ruleref uri="#Fbarra"/></item>
</rule>

<!-- sv -> sv,sp -->
<rule id="SV7">
  <item><ruleref special="GARBAGE"/></item>
  <!--<item><ruleref uri="#SV"/></item>-->
  <item><ruleref uri="#SP"/></item>
</rule>

<!-- sv -> sp,sv -->
<rule id="SV8">
  <item><ruleref uri="#SP"/></item>
  <item><ruleref uri="#SV"/></item>
</rule>

<rule id="det">
  <one-of>
    <item>a</item>
    <item>o</item>
    <item>os</item>
    <item>as</item>
    <item>algum</item>
  </one-of>
</rule>

<rule id="nome">
  <one-of>
    <item>acidente</item>
    <item>melhoras</item>
    <item>aguento</item>
    <item>antebraco</item>
    <item>braco</item>
    <item>carro</item>
    <item>causa</item>
    <item>coisas</item>
    <item>cotovelo</item>
    <item>dia</item>
    <item>dedos</item>
    <item>dor</item>
    <item>dores</item>
    <item>favor</item>
    <item>forca</item>
    <item>gelo</item>
    <item>lado</item>
    <item>maos</item>
    <item>noite</item>
    <item>nome</item>
    <item>novidade</item>
    <item>olhos</item>
```

```
    <item>parte</item>
    <item>partido</item>
    <item>radiografia</item>
    <item>problema</item>
    <item>prazer</item>
    <item>senhor</item>
    <item>senhora</item>
    <item>sensibilidade</item>
    <item>sitio</item>
    <item>tarde</item>
    <item>vida</item>
    <item>analgesico</item>
  </one-of>
</rule>

<rule id="verbo">
  <one-of>
    <item>passado</item>
    <item>gosta</item>
    <item>comer</item>
    <item>acho</item>
    <item>aconteceu</item>
    <item>ajudar</item>
    <item>aleijei</item>
    <item>aperte</item>
    <item>aleijou</item>
    <item>apresentar</item>
    <item>colocar</item>
    <item>conhecemos</item>
    <item>consegue</item>
    <item>consigo</item>
    <item>cortei</item>
    <item>deixe</item>
    <item>deve</item>
    <item>dobrar</item>
    <item>dobre</item>
    <item>doi</item>
    <item>doi-me</item>
    <item>estou</item>
    <item>esta</item>
    <item>estao</item>
    <item>faz</item>
    <item>fazer</item>
    <item>faca</item>
    <item>ficou</item>
    <item>foi</item>
    <item>gostaria</item>
    <item>magoei</item>
    <item>magoou</item>
    <item>manter</item>
    <item>mexa</item>
    <item>mexer</item>
    <item>mostre</item>
    <item>mover</item>
    <item>ocorreu</item>
    <item>passa</item>
```

```
    <item>percebi</item>
    <item>pode</item>
    <item>posso</item>
    <item>precisa</item>
    <item>prescrever</item>
    <item>ser</item>
    <item>sou</item>
    <item>sao</item>
    <item>tem</item>
    <item>tenha</item>
    <item>tenho</item>
    <item>tente</item>
    <item>ter</item>
    <item>tive</item>
    <item>traz</item>
    <item>vai</item>
    <item>vamos</item>
    <item>veio</item>
    <item>vejam</item>
    <item>vim</item>
    <item>vou</item>
    <item>vao</item>
    <item>confio</item>
    <item>irradia</item>
    <item>mexo</item>
    <item>relaxe</item>
    <item>repita</item>
  </one-of>
</rule>

<rule id="adjectivo">
  <one-of>
    <item>direito</item>
    <item>esquerdo</item>
    <item>esquerda</item>
    <item>belo</item>
    <item>boa</item>
    <item>bom</item>
    <item>bons</item>
    <item>identica</item>
    <item>igual</item>
    <item>necessario</item>
    <item>dolorido</item>
    <item>dorido</item>
    <item>imobilizado</item>
    <item>inchado</item>
    <item>magoado</item>
    <item>passado</item>
    <item>tratado</item>
  </one-of>
</rule>

</grammar>
```

Appendix D

Med-Speech transcription

Recordarei rapidamente as grandes linhas deste dossier.
Devemos exigir a reserva da menção ”chocolate de qualidade” aos produtos que contêm manteiga de cacau;
Subscrevo estas exigências que se encontram nas alterações apresentadas pelo relator.
Sou efectivamente um consumidor preocupado com a qualidade e o gosto do chocolate tradicional.
Muita parra pouca uva.
Não brinquem mais com a alimentação.
Isso sim isso seria um verdadeiro progresso.
Querem crer numa coisa destas?
Todos os gostos estão na natureza.
Os consumidores exigem uma informação precisa e clara.
Já não suportam ser enganados.
A base do bem estar na UE é a competitividade.
Não se trata de uma questão de saúde nem de protecção do meio ambiente.
Do ponto de vista dos consumidores é apenas uma questão de paladar.
Não poderíamos fazer outra coisa.
Os consumidores hão de saber o que sabe melhor!
Uma elaboração tão prolongada terá ao menos produzido um texto útil?
A resposta é claramente negativa.
A nova directiva não responde a qualquer necessidade.
Foram previstas algumas derrogações para levar em conta os hábitos alimentares de alguns países nórdicos.
A situação era pois satisfatória.
Porquê então mudá-la?
Porque é que sob pretexto de ”simplificá-la” se começou a subverter a directiva de mil novecentos e setenta e três?
O acervo comunitário terá deixado de ser intocável?

A manteiga de cacau representa efectivamente oito a nove porcento do preço total do chocolate.
De maneira nenhuma!
É exactamente o contrário O alinhamento processase como de costume pelo menor denominador comum.
O novo texto permitirá ao menos que o consumidor distinga facilmente entre chocolate verdadeiro e falso?
A Comissão está a fixar uma percentagem que sabe não ser verificável.
Chama-se a isto uma burla.
O texto inclui ainda disposições desonestas em matéria de rotulagem.
Uma vez que o Conselho não adoptou esta alteração o consumidor não será por isso correctamente informado.
Amantes do chocolate saboreiem religiosamente os vossos próximos ovos de Páscoa!
E depois do chocolate o que se segue?
Rotulagem quando nos tocas ...
Este é um símbolo distintivo a nível europeu ao qual queremos devolver o vigor.
Assenta no princípio de uma abordagem global que toma em consideração todo o ciclo de vida do produto.
Os símbolos nacionais não deverão no entanto ser todos colocados num plano secundário.
Quase que acreditaríamos estar perante uma barra de chocolate...
Como se tal não bastasse o Governo elevou recentemente aquele objectivo para sessenta setenta porcento até dois mil e cinquenta.
O Parlamento Europeu deveria incitar os outros Estados Membros a seguir o exemplo do Reino Unido.
Obviamente que a protecção do planeta implica muito mais do que a mera limpeza da atmosfera.
Implica proteger o ambiente natural igualmente á superfície da terra.
Também aqui o Governo Trabalhista britânico tomou a dianteira.
Os nossos cidadãos querem ver acções concretas.
Recomendação para segunda leitura Blokland
Já transmiti as minhas preocupações e as minhas interrogações quando da primeira leitura.
A maioria decidiu no sentido da coincineração o que não me satisfaz.
Todos sabem que o volume de resíduos a incinerar no seio da União Europeia irá aumentar.
O debate cristalizou-se sobretudo á volta das instalações de coincineração.
Recordo que estas instalações utilizam resíduos perigosos como combustíveis na produção de energia ou de elementos materiais.
é verdade que as disposições das directivas aplicáveis impõem uma limitação na utilização dos resíduos perigosos num nível máximo de quarenta porcento dos combustíveis utiliza-

dos. Esta percentagem é demasiado elevada!
Deverão aplicar-se normas de controlo particularmente rigorosas a estas instalações e estas normas deverão ser tão rigorosas como as aplicadas ás incineradoras.
Por outro lado insisto no facto de que esta generalização da exploração dos resíduos perigosos não deve levar a que nos acomodemos á produção de tais resíduos.
Com efeito corremos o risco enorme de amanhã ouvir dizer que se fizeram investimentos para construir estas coincineradoras e que é necessário rentabilizá-los!
Refuto de antemão vigorosamente estes argumentos que conduzirão a que o volume de transportes de matérias perigosas venha a aumentar com todos os riscos que tal situação implica para o meio ambiente e para a segurança.
Para concluir e relembrando que é necessário sancionar as infracções e ajudar as autarquias que pretendam normalizar as suas incineradoras há que insistir no facto de que a nossa prioridade é e deve continuar a ser a prevenção da produção de resíduos sejam estes perigosos ou não.
Nos últimos anos a União Europeia tem assumido a dianteira na concepção de novas estratégias em matéria de gestão de resíduos.
Assim a directiva que hoje nos ocupa tem por objectivo prevenir os efeitos negativos da incineração de resíduos a nível quer do ambiente quer da saúde humana mediante uma significativa redução dos níveis de emissão para a atmosfera de alguns importantes poluentes.
Esta directiva procura igualmente controlar a descarga de resíduos nas águas superficiais e subterrâneas.
Este relatório é-nos presente para a sua segunda leitura pelo Parlamento.
Considero tratar-se de uma proposta particularmente sensata.
é muito importante que a União Europeia se empenhe activamente na elaboração de novas propostas em matéria de gestão de resíduos.
A produção de resíduos na Europa processa-se a uma média de mais de 1 kg por pessoa por dia.
Todos os anos a UE gera mais de dois biliões de toneladas de resíduos.
Existe um aspecto da questão da eliminação de resíduos na União Europeia que não oferece quaisquer dúvidas.
Nomeadamente o de que o futuro da eliminação de resíduos não pode depender da busca de norte a sul e de leste a oeste por parte das autarquias locais de novos locais para a instalação de aterros sanitários.
Assiste-se a uma crescente preocupação por parte da opinião pública relativamente ás consequências dos aterros sanitários para o ambiente.
Impõe-se no mínimo uma mudança radical nesta área passando da dependência dos aterros sanitários para projectos integrados de gestão de resíduos.
A União Europeia apoia plenamente os objectivos estritos e ambiciosos mas nem por isso

inatingíveis traçados pelo Governo irlandês em matéria de reciclagem e recuperação.
Destinados a serem alcançados nos próximos quinze anos tais objectivos incluem nomeadamente o seguinte:
Um desvio para outros destinos de cinquenta porcento dos resíduos domésticos actualmente depositados em aterros sanitários.
Uma redução de sessenta e cinco porcento no mínimo dos resíduos biodegradáveis actualmente depositados em aterros sanitários.
O desenvolvimento de meios de recuperação de resíduos mediante o emprego de tecnologias compatíveis com o ambiente capazes de tratar anualmente até trezentas mil toneladas de resíduos biodegradáveis.
Todas estas medidas reflectem uma mudança a nível da opinião pública sobre a questão da gestão de resíduos.Os aterros sanitários têm os seus dias contados.
Impõe-se a sua substituição por tecnologias mais respeitadoras dos imperativos ambientais sendo esta uma abordagem que conta com o apoio unânime dos quinze Estados Membros da União Europeia.
Por este motivo o Grupo UEN não apoiou as alterações que são inaplicáveis nos nossos Estados e cujas consequências económicas previsíveis se revelariam muito nefastas.
Quanto ao resto e na minha qualidade de eleita local espero que uma fusão e uma simplificação das directivas comunitárias tornem mais claro e mais simples o quadro regulamentar no qual se devem inscrever os projectos municipais de gestão de resíduos. .
A adopção desta directiva relativa á incineração de resíduos que será aplicada a partir de dois mil e cinco melhora consideravelmente o quadro actual de gestão de resíduos na Europa promovendo critérios tão importantes como a saúde pública e a defesa dos direitos dos cidadãos.
Aliás se esta directiva entrasse de imediato em vigor o que infelizmente só acontece em dois mil e cinco seria praticamente impossível que o Governo português prosseguisse a sua estratégia de coincineração.
Em todo o caso sendo uma directiva que coloca elevadas restrições técnicas jurídicas e processuais a todos os processos de incineração e coincineração na Europa tornará mais evidente que a estratégia seguida em Portugal é profundamente errada.
Assim embora fosse meu desejo que se pudesse ter ido mais longe nas restrições a introduzir para os processos de incineração e de coincineração não posso deixar de me congratular com os resultados alcançados uma vez que esta directiva:
um Estabelece valores limite de emissões mais rígidos nomeadamente de óxidos de carbono e de poeiras.
dois Equipara ao nível dos valores limite as instalações de incineração e de coincineração acabando com o regime de excepção que estas detinham.
três Considera sem margem para dúvidas que a incineração e a coincineração só poderão ser utilizadas como um último recurso depois de esgotadas todas as outras abordagens

técnicas.
quatro Define condições muito rígidas ao nível da gestão global de resíduos da localização e dimensionamento das instalações de incineração e coincineração a serem cumpridas previamente á concessão de qualquer liderança.
cinco Aumenta o poder de intervenção das autoridades competentes face ao eventual incumprimento dos valores limite por parte das instalações.
seis Reduz drasticamente a possibilidade de os resíduos resultantes da coincineração e da incineração poderem ser utilizados nomeadamente na construção civil.
sete Introduz a amostragem contínua de dioxinas como um requisito obrigatório dos processos de incineração e de coincineração.
oito Aumenta a participação dos cidadãos no processo de decisão da localização das instalações.
nove Assegura a obrigatoriedade da publicação anual de um relatório ambiental que conclua sobre o cumprimento dos valores limite.
dez Limita consideravelmente as derrogações existentes para as instalações em fornos de cimento.
onze Define pela primeira vez um valorlimite para as emissões de amoníaco dada a acidificação que este composto químico produz na atmosfera.
No final da primeira leitura o Parlamento Europeu tinha imprimido claramente a sua marca nos debates ao propor a fusão de duas directivas sobre a incineração dos resíduos domésticos e a incineração dos resíduos perigosos.
Actualmente penso que a posição comum que nos é apresentada pelo Conselho constitui um compromisso muito bom que posso subscrever.
Todavia gostaria de abordar um ponto específico designadamente o da coincineração nas cimenteiras e justificar o meu voto negativo em relação a duas alterações.
Por um lado a primeira parte da alteração vinte cinco ao Anexo dois que fixa os valores limite das emissões para a indústria cimenteira e em particular o valor de quinze miligramas por metro quadrado para as poeiras.
A maior parte dos Estado Membros têm actualmente valores limite de cinquenta miligramas por metro quadrado.
Numa análise recente da relação custos/benefícios a Comissão concluiu que valores de trinta miligramas por metro quadrado seriam o melhor compromisso.
A imposição de normas de quinze miligramas por metro quadrado implicaria um custo ambiental excessivo e não seria rentável.
Por outro lado a alteração catorze sobre a definição dos resíduos urbanos reintroduz uma definição de resíduo urbano pouco clara quando esta definição já está prevista no número três do artigo terceiro.
Uma má interpretação deste artigo poderia levar á exclusão da utilização dos resíduos urbanos em fornos de cimenteiras. Ora este tipo de tratamento está previsto nos planos de

tratamento de resíduos de várias autoridades públicas.
Este voto não me coloca em contradição com as minhas preocupações ambientalistas pois a coincineração em cimenteiras não é especialmente nociva para o meio ambiente: os fornos de cimento estão efectivamente submetidos a regras rígidas de emissão.
Além disso a coincineração por vezes faz parte integrante da política de gestão dos resíduos de certos Estados ou de certas regiões em condições aceitáveis a nível económico e ambiental.
Tornar mais rigorosas as normas a que devem obedecer as unidades de incineração é altamente desejável.
No entanto é duvidoso que se deva estabelecer o mesmo tipo de normas para instalações concebidas para tratar resíduos diversos e para as que são destinadas a tratar apenas determinadas hierarquias de resíduos ou seja resíduos prétriados.
As instalações para resíduos não triados exigem equipamento mais avançado.
Fazer as mesmas exigências a todos os tipos de unidades de incineração pode ter como consequência desincentivar a triagem dos lixos e desse modo dificultar a reciclagem e a reutilização incluindo a compostagem dos resíduos orgânicos.
Assim nós democratascristãos suecos exortamos a Comissão a ter na devida conta estes pontos de vista no trabalho em curso relativo ás questões do lixo e dos resíduos.
O objectivo da directiva é louvável: trata-se de prevenir e reduzir os efeitos negativos da incineração e da coincineração dos resíduos.
Não podemos porém apoiar a proposta que visa submeter todas as instalações ás mesmas normas de emissão.
Não é por esta via indirecta que poderemos tratar a questão da sobrecapacidade das incineradoras na Europa.
O relator admite que certos fluxos de resíduos são mais fáceis de tratar no quadro de uma instalação de coincineração.
Pensamos que certas alterações propostas da posição comum correm de facto o risco de perturbar o nivelamento das instalações existentes e de penalizar a coincineração dos resíduos perigosos nas cimenteiras uma questão que diz essencialmente respeito á França e á Bélgica.
Relativamente aos resíduos urbanos gostaria de fazer várias observações.
Em primeiro lugar não é oportuno criar uma confusão em relação á definição dos ”resíduos urbanos mistos” no quadro da coincineração.
Equiparar os ”resíduos parcialmente separados” aos ”resíduos urbanos mistos não tratados” tal como se propõe na alteração catorze arrisca afectar o desenvolvimento das fileiras de triagem de resíduos que procuramos desenvolver e cujos resíduos ou certas fracções seriam susceptíveis de serem coincinerados.
Por outro lado se se alterasse o número quatro do artigo sétimo retirando desta forma a qualificação de ”não tratados” aos resíduos urbanos isso equivaleria a interditar a coincineração

de fracções de resíduos urbanos triados (papéis cartões...) que não poderiam ser reciclados excepto se fossem recolhidos separadamente.
Por último não é possível aceitar a redução do nível de seis toneladas por hora abaixo do qual as instalações existentes de incineração de lixos domésticos estão submetidas a um valorlimite de emissão atmosférica de duzentas miligramas por metro quadrado para os NOx pois a planificação pelas autarquias locais da adaptação dos equipamentos existentes opera-se com base neste nível (ver alínea a) do Anexo cinco).
Apesar de a incineração não constituir uma solução perfeita temos consciência disso a incineração representa um progresso importante para numerosas autarquias que aliás já autorizaram investimentos significativos para o desenvolvimento de uma gestão moderna do serviço municipal de tratamento de resíduos domésticos e similares.
é preciso levar isto em conta apesar de como é óbvio a reflexão e a investigação deverem prosseguir.
No futuro a indústria do tratamento de resíduos terá sem dúvida nenhuma que progredir muito para evitar ser ela própria uma poluidora.
Consequentemente deveremos elaborar estratégias de gestão dos resíduos de maior dimensão e incentivar activamente a investigação e o desenvolvimento nesta área essencial para a vida e o futuro das nossas autarquias locais.
Custa-me a acreditar atendendo á histeria anti Bruxelas dos seus aliados do UKIP que o senhor deputado Blokland tenha conseguido produzir tão monstruoso exemplo de excesso de regulamentação.
Resta-nos a satisfação de saber que as medidas propostas neste seu relatório irão permitir que o lixo anti europeu espalhado pelo senhor deputado Blokland e pelo UKIP será eliminado de uma forma respeitadora do ambiente.
Estas propostas têm por objectivo melhorar a qualidade do ar e proteger a saúde humana.
A incineração é não raras vezes o método mais eficaz e seguro em matéria de gestão de resíduos.
Por vezes porém pode libertar substâncias químicas perigosas para a atmosfera como é o caso das dioxinas.
é vital por conseguinte que estes produtos secundários sejam controlados e impedidos de escapar para o ambiente.
A tecnologia para limpeza de fumos emanados de instalações de incineração de resíduos já existe.
Cumpre-nos apenas assegurar que ela seja utilizada.
As medidas sensatas propostas pela Comissão são susceptíveis de encorajar a utilização deste tipo de tecnologia.
A hipocrisia do UKIP deixa-me estupefacto.
Desafio este partido a explicar aos seus apoiantes como é possível ser aliado de um tão óbvio apoiante do excesso de regulamentação comunitária.

Pretendendo embora ser inimigo dos eurocratas o UKIP é claramente um seu amigo.
Receio bem que a limpeza do ambiente acabe por ser canja comparada com a tarefa de persuadir o UKIP a limpar a sua própria casa.
Os prejudicados infelizmente serão os nossos próprios cidadãos.
A acreditar nos textos dos Antigos a eliminação dos resíduos já colocava problemas na Roma antiga mesmo que mais tarde os referidos resíduos tenham feito a felicidade de gerações de arqueólogos.
Na época a maior parte dos resíduos eram biodegradáveis o que está longe de ser o caso hoje em dia.
Já falámos do chocolate.
Pois bem eis uma maneira concreta de ilustrar a questão seguindo por exemplo o percurso de uma embalagem de uma barra de chocolate desde o caixote do lixo até á fábrica de incineração onde podemos descobrir que a combustão da folha de alumínio gera gases tóxicos.
Em matéria de gestão de resíduos confrontamo-nos com um desafio muito importante.
Penso no entanto que certos países ainda não estão preparados nem materialmente nem intelectualmente.
Hoje em dia um terço dos resíduos domésticos e similares tratados pelo parque francês de instalações de eliminação está entregue a trezentas incineradoras.
A incineração gera emissões poluentes: poeiras metais e dioxinas que temos de controlar através de um sistema de tratamento dos fumos.
A regulamentação em matéria de limitação dos produtos residuais tem vindo a ser progressivamente reforçada.
Evidentemente são requeridas acções para fazer respeitar as regras pois é necessário proteger a saúde e o meio ambiente mas não desejamos no que se refere á coincineração que o campo de aplicação e a definição das suas instalações sejam alterados.
Chamo a atenção dos meus colegas para um outro ponto que é certamente secundário.
No entanto a meu ver torna-se necessário um trabalho de comunicação na sequência da fusão destas directivas.
Efectivamente a população tem a impressão de que resíduos especiais perigosos não perigosos e urbanos vão ser todos misturados e sofrer um tratamento comum no mesmo forno e na mesma instalação.
Desta forma procederem eles próprios á triagem dos resíduos afigurar-se-lhes-ia inútil.
Como tal será necessário um verdadeiro trabalho de informação e de explicação.
Neste debate sobre a qualidade do ar a partir do relatório Davies estamos a discutir o problema da presença do ozono no ar ambiente.
Os efeitos do ozono na saúde são vários: irritação dos olhos do nariz e da garganta dificuldades respiratórias tosse dores de cabeça etc.
O ozono contribui igualmente para provocar crises de asma para deteriorar as funções

respiratórias e para aumentar a reactividade brônquica.
A longo prazo poderá mesmo provocar o envelhecimento prematuro dos pulmões.
Esta constatação é no mínimo aterradora e é por esta razão que acolho com a maior satisfação esta proposta de directiva!
A proposta fixa objectivos a longo prazo bem como valores alvo intermédios para o ozono no ar ambiente.
Introduz também um quadro de normas para a protecção da saúde humana e do ambiente.
Finalmente incita os Estados Membros a controlar as concentrações de ozono assim como a informar o público dos resultados das medições.
A directiva proporciona assim um termo de referência para os níveis de ozono no ar ambiente.
Para fixar os objectivos a longo prazo a Comissão propõe o recurso aos valores guia adoptados pela Organização Mundial de Saúde em mil novecentos e noventa e sete para a qualidade do ar.
O objectivo principal é impedir que se excedam estes limites evitando dessa forma os efeitos nocivos sobre a saúde humana e o ambiente.
Contudo as propostas da Comissão não fixam uma data para o cumprimentos destes requisitos o que é de lamentar.
São estabelecidos valores alvo como primeira etapa.
O objectivo é atingir estes limites tanto quanto possível até dois mil e dez.
Estes valores alvo baseiam-se também nos valores guia da OMS admitindo-se no entanto que possam ser excedidos num número limitado de dias por ano.
Gostaria de salientar que a proposta da Comissão não leva em linha de conta as modificações que poderão resultar dos compromissos assumidos pela Comunidade no que respeita ás alterações climáticas.
Ora esses compromissos prevêem uma redução de oito porcento nos gases causadores do efeito de estufa.
Se eles forem respeitados os EstadosMembros deverão ser mais ambiciosos no que respeita aos valoresalvo para a redução dos níveis de ozono.
Hoje debatemos uma proposta legislativa que nos diz directamente respeito pois visa melhorar a qualidade do ar ambiente.
Tratase de uma proposta de directiva sobre os valoreslimite nacionais de emissão de determinados poluentes atmosféricos (uma directiva que emana da directivaquadro sobre a qualidade do ar).
Inscrevese no Quinto Programa de Acção da Comunidade em matéria de Ambiente.
Os valores limite indicados neste texto devem ser respeitados o mais tardar até dois mil e dez.
Nesta perspectiva os EstadosMembros deverão elaborar programas para a redução progressiva das suas emissões nacionais anuais e transmitilos á Comissão antes do final de

dois mil e dois.
Os programas deverão ser revistos e actualizados até dois mil e seis.
Estes inventários e previsões deverão ser transmitidos anualmente á Comissão.
A senhora relatora propôs alterações com o objectivo de reforçar esta proposta.
Sugere nomeadamente a fixação de normas mais rígidas que deverão ser respeitadas em termos de emissões dos quatro poluentes em causa.
Quanto ao princípio não posso deixar de aderir a estas propostas.
Efectivamente quando se trata da saúde e do ambiente é necessário revelarmonos ambiciosos se pretendemos melhorar a situação.
No entanto colocar a fasquia demasiado alta pode revelar-se contraproducente.
Por este motivo penso que de momento mais vale fixarmonos nos objectivos estabelecidos nas convenções internacionais sabendo que será sempre possível rever os valoreslimite perante os progressos já realizados.
De resto o relatório solicita á Comissão que proponha novas alterações á directiva até dois mil e quatro.
Deste modo a Comissão poderá rever os valores limite fixados no texto os objectivos ambientais intermédios e as novas medidas tomadas para assegurar o respeito dos valores limite.
Essas eventuais correcções eventuais permitirão adaptar a directiva aos progressos técnicos e científicos e eventualmente atingir os objectivos estabelecidos para dois mil e dez.
Do meu ponto de vista tratase de um bom compromisso.
A senhora relatora propõe também a fixação de datasalvo para a concretização dos objectivos fixados; isto pareceme indispensável.
Efectivamente sem prazos a cumprir os textos não passam com demasiada frequência de declarações de intenção sem seguimento.
No mesmo sentido deverá ser implementado o mais rapidamente possível um dispositivo de sanções aplicáveis ás violações das disposições da directiva.
Somos a favor de que a UE através de uma legislaçãoquadro estabeleça valoreslimite no domínio do ambiente uma vez que a maior parte dos problemas ambientais são generalizados e ultrapassam fronteiras.
Proposta de resolução sobre o Conselho Europeu de Lisboa
A ambição do pleno emprego uma expressão que voltou á ordem do dia com o retorno de um certo crescimento é certamente uma intenção louvável; no entanto se não atacarmos as causas reais do desemprego será difícil atingila.
é do conhecimento de todos nós que apesar do crescimento a criação de empregos na Europa é manifestamente insuficiente.
Os empregos criados caracterizamse pela precariedade nomeadamente no que se refere ás mulheres e aos jovens.
Temos que deixarnos de declarações de intenções por mais positivas que elas sejam.

Face a expectativas sociais tão fortes é urgente definir objectivos mensuráveis e verificáveis. Gostaria muito que tivesse sido assumido um compromisso em relação ás propostas formuladas na resolução do meu grupo por exemplo através:
Do lançamento de um processo de redução do horário de trabalho sem quebra salarial nem flexibilidade;
De uma revisão das directivas sobre os comités de empresa europeus e os despedimentos colectivos no sentido de reforçar os direitos e os poderes dos assalariados e das suas organizações sindicais nomeadamente nos casos de fusão e de restruturação.

Appendix E

Technology Survey

The next tables comprise the technology survey of the components for the development of the S2ST system. The prices are in Euros.

System	Dev. / Vendors	Last Release	Begin	Active	Type	Price	Users	EN	PT	SDK
LumenVox	LumenVox	9.5	2001	yes	commercial	2,800	business	yes	no	yes
Dragon Professional	Nuance	12	1992	yes	commercial	480	individual/business	yes	yes	yes
Dragon Medical 360	Nuance	?	?	yes	commercial	?	business	yes	?	?
Hidden Markov Toolkit	CUED/MS	3.4.1	1993	?	open		academic research	no	no	yes
Julius	Nagoya IT	4.2.2	1997	yes	open		academic research	yes	no	yes
CMU Sphinx	CMU	4	2000	yes	open		academic research	yes	no	yes
Google Voice	Google	?	2011	yes	free		individual/developers	yes	yes	no
Microsoft Speech P.	Microsoft	11	?	yes	free		developers/business	yes	yes	yes

Table E.1: Survey of ASR systems

System	Dev. / Vendors	Approach	Last Release	Begin	Active	Type	Price	Users	EN	PT
SYSTRAN	SYSTRAN	hybrid	?	1968	yes	commercial	900	business	yes	yes
PROMT	PROMT	rule-based	?	1991	yes	commercial	39-1,200	business	yes	yes
SDL Lang. Weaver	SDL	statistical	?	2002	yes	commercial	100-2,500	business	yes	yes
Google Translate	Google	statistical	?	2006	yes	commercial	?	individuals/business	yes	yes
Bing	Microsoft	statistical	?	2009	yes	commercial	30-4,500	ind./dev./business	yes	yes
Websphere Trans.	IBM	?	?	1998	yes	commercial	20,000	business	yes	yes
Apertium	Alacant Univ.	rule-based	3.2	2004	yes	open		academic research	yes	yes
Moses	Moses	statistical	1	?	yes	open		ind./dev./business	yes	yes
NiuTrans	Northeastern Univ.	hybrid	1.2.0	2011	yes	open		academic research	yes	yes

Table E.2: Survey of MT systems

System	Dev. / Vendors	Approach	Last Release	Begin	Active	Type	Price	Users	EN	PT
Loquendo	Nuance	?	?	2000	yes	commercial	45	individuals	yes	yes
Cepstral	Cepstral LLC	?	?	2005	yes	commercial	155	ind./business	yes	no
Ivona	Amazon	concatenative	?	2003	yes	commercial	825	ind./dev./business	yes	no
Acapela	Acapela Group	?	?	1998	yes	commercial	?	ind./dev./business	yes	yes
Natural Voices	AT&T	?	?	1995	yes	commercial	800	business	yes	no
eSpeak	Jonathan Dudd.	formant	1.47.11	?	yes	open		academic research	yes	yes
Praat	Amsterdam Univ.	formant	5.3.52	?	yes	open		academic research	yes	no
Festival	Edinburgh Univ.	concatenative	2.1	2006	?	open		academic research	yes	no
MARY	Saarland Univ.	concatenative	5	?	yes	open		academic research	yes	no
Microsoft SP	Microsoft	concatenative	11	?	yes	free		dev./business	yes	yes

Table E.3: Survey of TTS systems

Appendix F

Resources and tools used

Documentation

Various tools were used to implement the S2ST prototype. In several occasions I used the documentation that came available with the tools. The following items reference these along with the link (active as of 18-11-2013) from where they can be obtained:

- Moses User Manual and Code Guide: Introduces the technology behind the Moses statistical machine translation system, explains how to install the system and provides tutorials and guidelines for the training of a SMT system. The June 25, 2012 version was used. Available at `http://www.statmt.org/moses/manual/manual.pdf`
- MSDN: The Microsoft Developer Network provides the information and documentation for the implementation of the ASR and TTS system as also the reference for the programming in Visual Studio of the web services and set up of the IIS server. The MSDN is available at `http://msdn.microsoft.com/en-US/`, the manuals used were the Microsoft Speech Platform SDK 11 Documentation, available at `http://msdn.microsoft.com/en-us/library/dd266409(v=office.14).aspx` and the web services manuals at `http://msdn.microsoft.com/en-us/library/ms950421.aspx`, the documentation is not versioned.
- iOS Developer Library: Apple provides, through this on-line library, the necessary documentation and manuals for the development of an IOS Application, `https://developer.apple.com/library/ios/navigation/`. The manuals used were the iOS Human Interface Guidelines (2012-03-07 version) , the Objective-C Programming Language Manual ((2011-10-12 version) and the Object-Oriented Programming with Objective-C (2010-11-15 version).
- SRGS: Speech Recognition Grammar Specification Version 1.0 documents were used for the implementation of the SRGS grammar. It is available at `http://www.w3.org/TR/speech-grammar/`

- Web Speech API Specification: For the implementation of the web based prototype the Web Speech API Specification (2012-10-19 version) was used in regard to the ASR component. It is available at `https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html`

  All the documentation that may be distributed is available at `http://nlx.di.fc.ul.pt/~s2st/resources/`.

Resources

The following items list the parallel corpora, with Portuguese and English language, that was used in this work, all of the links were active as of 18-11-2013.

- Europarl: The version 7 was used and it is available at `http://www.statmt.org/europarl/`
- Jrc-Acquis: The version 3 was used and it is available at `http://ipsc.jrc.ec.europa.eu/index.php?id=198`
- EMEA: The version 0.3 was used and it is available at `http://www.ema.europa.eu/ema/`
- ECB: The version 0.1 was used and it is available at `http://opus.lingfil.uu.se/ECB.php`
- ACL 2008: The development, test and training parallel corpora were used. Available at `http://www.statmt.org/wmt08/shared-task.html`
- EUbookshop: The version compiled in the year 2012 was used and it is available at `http://opus.lingfil.uu.se/EUbookshop.php`
- KDE4: The version 2 was used and it is available at `http://opus.lingfil.uu.se/KDE4.php`
- Tatoeba: The version compiled in the year 2012 was used and it is available at `http://opus.lingfil.uu.se/Tatoeba.php`
- EUconst: The version compiled in the year 2009 was used and it is available at `http://opus.lingfil.uu.se/EUconst.php`
- UMLS: Although the Unified Medical Language System is a tool used for interface with medical information, its database was used for the extraction of medical corpora. The databases used were: UMLS (2013AA version), RxNorm (2005 version), SNOMED CT (2007 version) and CMT. It is available at `http://www.nlm.nih.gov/research/umls/`

For the ASR component the Microsoft Speech Platform Server Runtime Languages were used (English and Portuguese, version 10.2). These are available at http://www.microsoft.com/en-us/download/details.aspx?id=21924

All the parallel corpora are also available at http://nlx.di.fc.ul.pt/~s2st/resources/.

Tools

The following items list the tools used for the implementation of the S2ST prototype, all of the links were active as of 18-11-2013.

- Moses: A Statistical Machine Translation System used for the implementation of the SMT component of the S2ST prototype. It provides the backbone for the training and evaluation of a SMT system with a combination of different sub-tools. The version 1.0 was used and it is available at http://www.statmt.org/moses/?n=Development.GetStarted

- LX Suite: The Language Resources and Technology for Portuguese, from the NLX-Natural Language and Speech Group at the University of Lisbon, Department of Informatics, is a set of tools that provide natural language processing for the Portuguese language. It was used for the tokenization, part-of-speech tagging and lemmatizing of the corpora used for the training of the SMT. These set of tools are available at http://lxcenter.di.fc.ul.pt/index.html.

- Stanford CoreNLP: A set of tools developed by the Stanford NLP Group that provide the natural language processing for the English language. It was used for the tokenization, part-of-speech tagging and lemmatizing of the corpora used for the training of the SMT. The version 3.2.0 was used and it is available at http://nlp.stanford.edu/software/corenlp.shtml

- Gizza++: A statistical machine translation toolkit that is used to train IBM Models 1-5 and an HMM word alignment model. It was used in conjunction with Moses for the training of the SMT component. The version 1.0.7 was used and it is available at http://code.google.com/p/giza-pp/downloads/list

- SRILM: The SRI Language Modeling Toolkit is a toolkit for building and applying statistical language models (LMs). It was used for the creation of the language models used in the ASR, SMT and TTS component. The 1.7.0 version was used and it is available at http://www.speech.sri.com/projects/srilm/download.html

- Apertium toolbox: It was used a script, apertium-eval-translator, written in Perl that allows the calculation of the word error rate. It was used for the evaluation of the ASR word error rate. The 48725 revision was used. Available at `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-eval-translator/`

- Microsoft Visual Studio and IIS: These tools provided the development and runtime environment for the ASR and TTS component along with their web services. The Visual Studio 2010 Ultimate version was used and the IIS version 7. These tools are available at `http://www.microsoft.com/visualstudio/ptb/products/visual-studio-2010-express` and `http://www.microsoft.com/en-us/download/details.aspx?id=2299`

- Microsoft Speech SDK: A set of development tools for building voice-enabled applications that leverage Microsoft's redistributable speech engines. The version 11 was used for the implementation of the ASR and TTS component. It is available at `http://www.microsoft.com/en-us/download/details.aspx?id=27226`

- XCode and iOS SDK: An integrated development environment containing a suite of software development tools by Apple for developing software for OS X and a SDK containing the necessary tools for the development of the mobile application. The XCode used was the 4.2 version. The XCode is available at `https://developer.apple.com/xcode/` and the iOS SDK is downloaded from within the XCode preferences.

Appendix G

Resources developed

The following items list the resources developed for the implementation of the S2ST prototype, all of these are available at `http://nlx.di.fc.ul.pt/~s2st/resources/`.

- Anamnesis Corpus: Hypothetical doctor-patient dialogue, in Portuguese, to address an arm injury doctor-patient diagnosis and treatment corpora (vd. Appendix A). The assembled dialogue has 160 sentences and contains 445 tokens. The dialogue methodology and typical dialogue context were done with the help of an experienced family physician, with the application of the anamnesis process.

- Med-CFG: A context free grammar created manually and solely from the Anamnesis Corpus (vd. Appendix B).

- Med-SRGS: A SRGS grammar created manually from the context free grammar rules (vd. Appendix C).

- Med-Speech: A speech corpus with the recording voice of 200 portuguese sentences used for the evaluation of the ASR component. The Med-Speech Corpus transcription consists of 200 spoken sentences, with a total of 3911 words, with approximately 20 words per sentence (vd. Appendix D).

- EUMed Corpora and EUMed Corpora Annotated: The collection of diverse parallel corpus and the resulting corpora with the annotation provided by the LX-Suite and CoreNLP, containing surface, lemma and part of speech forms, used for the training of the SMT component (vd. section 4.1). The corpora used consists of: the Europarl (1,960,407 sentences for each language), JRC-Acquis (1,236,846 sentences for each language), UMLS (85,501 sentences for each language), EMEA (1,082,144 sentences for each language) and ECB (202,030 sentences for each language).

Bibliography

[1] X. Huang and L. Deng, *Handbook of Natural Language Processing*, ch. An overview of modern speech recognition, pp. 339–366. Boca Raton, FL, USA: CRC, Taylor and Francis, 2010.

[2] M. P. e. Lewis, *Ethnologue: Languages of the World*. Dallas, Texas: SIL International, sixteenth ed., 2009.

[3] P. F. Brown, J. Cocke, S. Della Pietra, and V. J. Della Pietra, "A statistical approach to french/english translation," in *Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI)*, 1988.

[4] P. Koehn, *Statistical Machine Translation*. New York, Cambridge University Press, 2010.

[5] R. Mitkov, *The Oxford Handbook of Computational Linguistics*. New York, Oxford University Press Inc., 2004.

[6] J. Jurafsky, Daniel. H. Martin, *Speech and Language processing*. New Jersey, Pearson Education Inc., 2009.

[7] W. J. Hutchins, "Machine translation: A concise history," in *Computer aided translation: Theory and practice*, Chinese University of Hong Kong, 2007.

[8] M. Rayner, P. Bouillon, B. Hockey, N. Chatzichrisafis, M. Starlander, *et al.*, "Comparing rule-based and statistical approaches to speech understanding in a limited domain speech translation system," in *10th International Conference on Theoretical and Methodological Issues in Machine Translation*, 2004.

[9] G. Thurmair, "Comparing rule-based and statistical machine translation output," in *Workshop on the amazing utility of parallel and comparable corpora, LREC*, p. 5, 2004.

[10] P. Koehn and H. Hoang, "Factored translation models," in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, vol. 868, p. 876, 2007.

[11] P. Brown, V. Pietra, S. Pietra, and R. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[12] A. Berger, V. Pietra, and S. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[13] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL, (Stroudsburg, PA, USA), pp. 48–54, Association for Computational Linguistics, 2003.

[14] D. Chiang, "A hierarchical phrase-based model for statistical machine translation," in *43rd Annual Meeting on Association for Computational Linguistics*, pp. 263–270, Association for Computational Linguistics, 2005.

[15] D. Chiang, "Hierarchical phrase-based translation," *Computational linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[16] K. Knight and D. Marcu, "Machine translation in the year 2004," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. v–965, IEEE, 2005.

[17] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, "Large language models in machine translation," in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, (Prague, Czech Republic), pp. 858–867, Association for Computational Linguistics, June 2007.

[18] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT Summit*, vol. 5, 2005.

[19] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga, "The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages," in *5th International Conference on Language Resources and Evaluation (LREC)*, 2006.

[20] D. Munteanu and D. Marcu, "Improving machine translation performance by exploiting non-parallel corpora," *Computational Linguistics*, vol. 31, no. 4, pp. 477–504, 2005.

[21] B. J. Dorr, E. H. Hovy, and L. S. Levin, *Encyclopedia of Language and Linguistics*. Elsevier, 2004.

[22] L. Deng and X. Huang, "Challenges in adopting speech recognition," *Commun. ACM*, vol. 47, pp. 69–75, Jan. 2004.

[23] S. Gaikwad, B. Gawali, and P. Yannawar, "A review on speech recognition technique," *International Journal of Computer Applications IJCA*, vol. 10, no. 3, pp. 24–28, 2010.

[24] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, 2012.

[25] R. Cole, J. Mariani, H. Uszkoreit, G. Varile, A. Zaenen, and A. Zampolli, *Survey of the state of the art in human language technology*, vol. 12. Cambridge University Press, 1998.

[26] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[27] R. Sproat, A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of non-standard words," *Computer Speech & Language*, vol. 15, no. 3, pp. 287–333, 2001.

[28] D. Yarowsky, "Hierarchical decision lists for word sense disambiguation," *Computers and the Humanities*, vol. 34, no. 1, pp. 179–186, 2000.

[29] S. Jiampojamarn and G. Kondrak, "Online discriminative training for grapheme-to-phoneme conversion," in *Interspeech*, pp. 1303–1306, Citeseer, 2009.

[30] H. M. Meng, "Phonological parsing for bi-directional letter-to-sound/sound-to-letter generation," tech. rep., Massachusetts Institute of Technology, Cambridge, MA, USA, 1995.

[31] S. Chen, "Conditional and joint models for grapheme-to-phoneme conversion," in *EUROSPEECH*, pp. 2033–2036, 2003.

[32] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[33] P. Lehnen, S. Hahn, A. Guta, and H. Ney, "Incorporating alignments into conditional random fields for grapheme to phoneme conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4916–4919, IEEE, 2011.

[34] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96)*, vol. 1, pp. 373–376, IEEE, 1996.

[35] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[36] C. Hamon, E. Mouline, and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modifications of speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 238 –241, 1989.

[37] K. Tokuda, H. Zen, and A. Black, "An hmm-based speech synthesis system applied to english," in *IEEE Workshop on Speech Synthesis*, pp. 227–230, IEEE, 2002.

[38] H. Zen and K. Tokuda, "Techware: Hmm-based speech synthesis resources," *Signal Processing Magazine, IEEE*, vol. 26, no. 4, pp. 95–97, 2009.

[39] T. Schultz and K. Kirchhoff, *Multilingual speech processing*. Academic Press, 2006.

[40] K. Hashimoto, J. Yamagishi, W. Byrne, S. King, and K. Tokuda, "Impacts of machine translation and speech synthesis on speech-to-speech translation," *Speech Communication*, 2012.

[41] Y. Liu, M. Harper, M. Johnson, and L. Jamieson, "The effect of pruning and compression on graphical representations of the output of a speech recognizer," *Computer Speech & Language*, vol. 17, no. 4, pp. 329–356, 2003.

[42] S. Rao, I. Lane, and T. Schultz, "Improving spoken language translation by automatic disfluency removal: Evidence from conversational speech transcripts," *Training*, vol. 6370, no. 46300, pp. 6–50, 2007.

[43] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper, "Comparing hmm, maximum entropy, and conditional random fields for disfluency detection," in *Eurospeach Conference on Speech Communication and Technology (INTERSPEECH)*, Citeseer, 2005.

[44] A. Parlikar, A. Black, and S. Vogel, "Improving speech synthesis of machine translation output," in *Interspeech*, pp. 194–197, 2010.

[45] I. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, "On the use of information retrieval measures for speech recognition evaluation," *Institut Dalle Molle Intelligence Artificielle Perceptive (IDIAP)*, 2005.

[46] A. Black and K. Tokuda, "The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common databases," in *Interspeech*, 2005.

[47] *MST: Proceedings of the Workshop on Medical Speech Translation*, (Stroudsburg, PA, USA), Association for Computational Linguistics, 2006.

[48] H. Somers, "Medical spoken language translation: What do the users really need?," in *International Conference on Translating and the Computer*, 2007.

[49] S. Narayanan, S. Ananthakrishnan, R. Belvin, E. Ettelaie, S. Gandhe, S. Ganjavi, P. Georgiou, C. Hein, S. Kadambe, K. Knight, *et al.*, "The transonics spoken dialogue translator: an aid for english-persian doctor-patient interviews," in *Working Notes of the AAAI Fall symposium on Dialogue Systems for Health Communication*, pp. 97–103, 2004.

[50] P. Bouillon, G. Flores, M. Georgescul, S. Halimi, B. Hockey, H. Isahara, K. Kanzaki, Y. Nakao, M. Rayner, M. Santaholma, *et al.*, "Many-to-many multilingual medical speech translation on a pda," in *Conference of the Association for Machine Translation in the Americas*, 2008.

[51] Y. Gao, L. Gu, B. Zhou, R. Sarikaya, M. Afify, H. Kuo, W. Zhu, Y. Deng, C. Prosser, W. Zhang, *et al.*, "Ibm mastor system: Multilingual automatic speech-to-speech translator," in *Workshop on Medical Speech Translation*, pp. 53–56, Association for Computational Linguistics, 2006.

[52] E. Ettelaie, S. Gandhe, P. Georgiou, K. Knight, D. Marcu, S. Narayanan, D. Traum, and R. Belvin, "Transonics: A practical speech-to-speech translator for english-farsi medical dialogues," in *Association for Computational Linguistics (ACL)*, pp. 89–92, Association for Computational Linguistics, 2005.

[53] M. Rayner, P. Bouillon, J. Brotanek, G. Flores, S. Halimi, B. A. Hockey, H. Isahara, K. Kanzaki, E. Kron, Y. Nakao, *et al.*, "The 2008 medslt system," in *International Conference on Computational Linguistics*, p. 32, 2008.

[54] K. E. Campbell, D. E. Oliver, and E. H. Shortliffe, "The unified medical language system toward a collaborative approach for solving terminologic problems," *Journal of the American Medical Informatics Association*, vol. 5, no. 1, pp. 12–16, 1998.

[55] J. Tiedemann, L. Nygaard, and T. Hf, "The opus corpus – parallel and free," in *4th International Conference on Language Resources and Evaluation (LREC)*, pp. 1183–1186, 2004.

[56] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, pp. 252–259, 2003.

[57] A. Branco and J. R. Silva, "A suite of shallow processing tools for portuguese: Lx-suite," in *European Chapter of the Association for Computational Linguistics*, pp. 179–182, 2006.

[58] C. Callison-Burch, P. Koehn, C. Monz, J. Schroeder, and C. S. Fordyce, eds., *Proceedings of the Third Workshop on Statistical Machine Translation*, (Columbus, Ohio), Association for Computational Linguistics, June 2008.

[59] M. F. M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the iwslt 2012 evaluation campaign," in *International Workshop on Spoken Language Translation (IWSLT)*, 2012.

[60] H. Yamamoto, Y. Wu, C.-L. Huang, X. Lu, P. R. Dixon, S. Matsuda, C. Hori, and H. Kashioka, "The nict asr system for iwslt 2012," in *International Workshop on Spoken Language Translation (IWSLT)*, 2012.

[61] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "Further meta-evaluation of machine translation," in *Third Workshop on Statistical Machine Translation*, (Columbus, Ohio), pp. 70–106, Association for Computational Linguistics, June 2008.

[62] N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, eds., *Parallel Data, Tools and Interfaces in OPUS*, European Language Resources Association (ELRA), 2012.